# STATISTICS

BY

WILLIAM VERNON LOVITT, Ph.D.
PROFESSOR OF MATHEMATICS, COLORADO COLLEGE

AND

HENRY F. HOLTZCLAW, Ph.D.
PROFESSOR OF COMMERCE, UNIVERSITY OF KANSAS

New York
PRENTICE-HALL, INC.
1931

PRINTED IN THE UNITED STATES OF AMERICA

# PREFACE

This book has developed over a period of years during the teaching of a first course in Statistics. The illustrative examples are the ones found best adapted to classroom use. Many examples from everyday life are included. The exercises are those which have appealed most to the interest of the students.

Too often, in the world-at-large, things easy are made difficult through an unwise choice and arrangement of words. Therefore, an attempt has been made throughout to put the discussion in the simplest possible words and in a non-technical form.

Numerous exercises have been placed at convenient intervals. In many of the lists some additional exercises are given where the data are not included. However, exact references are given as to where to find data suitable for use in connection with the theory under discussion. This obviates the necessity of the teacher or student losing valuable time in a search for suitable laboratory material.

For most of the text, the only mathematical knowledge required is some facility in numerical computations. For a small part, the equivalent of one year of algebra is required. An attempt has been made to keep the mathematics as simple as possible and at the same time not omit any essential part of the elementary theory of statistics. Some of the more difficult mathematical passages may be omitted without destroying the continuity of the course.

Data for both the illustrative examples and the exercises are drawn from a variety of sources. Perhaps the greatest amount is from the field of economics. Other fields from which data are taken are agriculture, biology, botany, business, education, medicine, and sociology.

Answers are given to many of the exercises.

The chapters on Graphical Representation, Index Numbers, and Correlation cover more ground than is usual in an introductory text on Statistics.

W. V. L.
H. F. H.

# CONTENTS

B. Seasonal Variation

C. Long Time Cycles

D. Lead and Lag

E. Babsonchart

Determination of the Parameters

## APPENDIX

# STATISTICS

## Chapter I

## INTRODUCTION

**1. Meaning of statistics.**—Statistics is the science which deals with the collection, classification, and tabulation of numerical facts, as a basis for the explanation, description, and comparison of phenomena. G. U. Yule in his "Introduction to the Theory of Statistics," p. 5, says: "By statistical methods we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes. By theory of statistics we mean the exposition of statistical methods."

The total population at any one time is affected by a multiplicity of causes. Among these causes are emigration, immigration, death rate, and birth rate.

Table 1, showing the total population of the United States for the years indicated, contains statistical facts which have been collected, classified, and tabulated by the United States Census Bureau. Tables of this kind are useful in estimating the population of the United States for the intervening years, and in computing the rate of growth of the population for specified periods. Also, the rate of growth of population in the United States may be compared with the rate of growth of population in some other country for corresponding periods.

**Table 1.—Population of the United States.**

| *Census Year* | *Total Population* |
|---|---|
| 1920 | 105,710,620 |
| 1910 | 91,972,266 |
| 1900 | 75,994,575 |

*Fourteenth Census of the U. S., 1920, Vol. II, Population, p. 29.*

In general, in statistics, we are interested not only in facts but in the methods by which facts may be secured and estimates and comparisons made. We are also interested in the principles—mathematical, economic, or otherwise—which justify these methods.

**2. Series.**—Usually the statistical facts, or data, are presented in the form of a table, as in table 1. The data are presented in two parallel columns. The first column gives the essential facts for distinguishing among the different numbers in the second column. The data in the first column give the scheme whereby the data in the second column are arranged into groups. The scheme is designated as a *rule of classification.* Each number in the second column is a *statistical item.* A set of statistical items, arranged with respect to a definite rule of classification, is a *statistical series.*

**3. Types of series**—(a) *Time series.*—Table 1 is representative of one type of series which we shall designate as a *time series.* The statistical items in column two are arranged in chronological order, i.e., in historical order. This order should be used only in those cases in which time is significant.

(b) *Frequency series.*—A set of things under observation may vary in the size or amount of the characteristic under observation. The statistical items in the second column

**Table 2.—Bituminous Coal Mines, United States, 1924.**

| *Classes* | *No. of Laborers outside the Mine* |
|---|---|
| under 30c | 417 |
| 30 and under 40 | 1,141 |
| 40 " " 50 | 1,707 |
| 50 " " 60 | 1,244 |
| 60 " " 70 | 304 |
| 70 " " 80 | 322 |
| 80 " " 90 | 2,042 |
| 90 " " 1.00 | 330 |
| 1.00 " " 1.10 | 7 |
| | TOTAL 7,514 |

*Bulletin No. 416, U. S. Bureau of Labor Statistics, p. 59.*

(see table 2) represent then the relative frequency of occurrence of the size or amount specified in the first column. Such a series is called a *frequency series* and the corresponding table a *frequency table.* This type of series is illustrated in table 2.

(c) *Categorical series.*—In a *categorical series* the items refer to a set of distinct things. The items of a categorical series are classified by a division of the items under consideration into certain broad categories, each category possessing a distinctive attribute. For example, the classification may be with respect to geographical location, nationality, name of firm, name of commodity, kind of industry, kind of occupation, etc. This type of series is illustrated by table 3.

**Table 3.—Sheep, Including Lambs, on Farms, in Thousands, Jan. 1, 1925.**

| STATE....... | *Texas* | *Iowa* | *Nebraska* | *Montana* | *Ohio* | *Illinois* |
|---|---|---|---|---|---|---|
| NUMBER..... | 3,465 | 891 | 840 | 2,579 | 2,178 | 694 |

*Yearbook of U. S. Dep't of Agriculture, 1925, p. 1148.*

**4. Variable.**—Any quantity which changes in magnitude is called a *variable.* The thing observed is a *variate.* The observed values are called *variate values.* Examples of a variable are: the population of the United States; the amount of coal mined in Pennsylvania; the amount of wheat harvested in Kansas; the number of automobiles sold in Colorado; the capacity of freight cars; and the price of vegetables at a nearby produce market.

Variables are of two kinds: continuous and discontinuous. A continuous variable is one which may take on all possible values between certain limits. These values are only approximations to the absolute values. They do not represent complete accuracy. A specific value represents only a range between certain limits. A measurement is never so exact but that further refinement is possible. The height of individuals is a continuous variable and may be measured to the nearest quarter-inch. Obviously, measurements of individuals could be recorded to a greater degree of accuracy; the degree depending to a large extent upon the effort put forth and the instruments available for measurement. Other examples of a continuous variable are: the age of an individual; the lengths and weights of ears of corn; the distance traveled by an automobile; the temperature at the summit of Pike's Peak; and the barometric pressure at sea level in New York City.

A discontinuous, or discrete, variable is one having breaks or gaps in the value of the variable. This is true regardless of the number of cases. Discrete data are determined with greater accuracy than continuous data. In a shoe factory there may be employed 99, 100, or 101 men, but it is never possible to employ $99\frac{1}{2}$ men or $100\frac{3}{4}$ men. In other words, the data are recorded with complete accuracy at definite positions in the interval within which no values can occur.

A discrete series arises whenever the variate values are obtained by counting and not by measurement. Data obtained in many economic and social investigations give rise to discrete series. Stock, quoted in minimum units of one-eighth of a dollar, and population statistics, given in whole numbers, are examples of discrete series.

**5. Class interval and class limits.**—The total number of observed values of a variate may be divided into groups called *classes.* The number of observations in each class is called the *class frequency.* The range of values for the variates in any given class is the *class interval* for that class. The pairs of numbers given in the column of classes (the first column) are the lower and upper *class limits.* The number halfway between the upper limit of one class and the lower limit of the next higher class is called a *class boundary.* The *upper boundary* of one class is the *lower boundary* of the next class. The number halfway between the upper and lower limits of a class is sometimes referred to as the *class mark.*

For example, in table 2, we have for the class interval ten cents. The frequency of the first class is 417. For the second class the lower limit is 30¢ and the upper limit is 40¢. The upper boundary of the second class, namely

40¢, coincides with the lower boundary, namely 40¢, of the third class. The class mark for the second class is 35¢.

In table 4 the class interval is 15. For the second class the class limits are 15 and 29. The class boundaries of the second class are $14\frac{1}{2}$ and $29\frac{1}{2}$. The class mark for the second class is 22.

**Table 4.—A Biometrical Study of Egg Production.**

$\overline{X} = 113.44$; $\sigma = 35.48$

| *No. of Eggs* | *Frequency* | *No. of Eggs* | *Frequency* |
|---|---|---|---|
| 0– 14 | 2 | 105–119 | 39 |
| 15– 29 | 2 | 120–134 | 26 |
| 30– 44 | 5 | 135–149 | 21 |
| 45– 59 | 5 | 150–164 | 19 |
| 60– 74 | 9 | 165–179 | 12 |
| 75– 89 | 16 | 180–194 | 1 |
| 90–104 | 30 | | |
| | | | Total 187 |

"*Domestic Fowl,*" *Bulletin No. 110, Bureau of Animal Husbandry, U. S. Dep't of Agriculture.*

### Exercises.

1. Construct tables illustrating time series, frequency series, categorical series.

2. Construct tables illustrating discrete series and continuous series.

3. In each series presented in Ex. 1 and Ex. 2 give the class interval, class limits, class boundaries, class marks.

## Chapter II

## PRIMARY AND SECONDARY DATA

**6. Definitions.**—All data are either primary or secondary. Data are said to be primary when collected by one's self for one's own use or when collected by someone else under one's direction for one's own use. The term "primary data" refers to the material as collected, to the original data before they have been sorted and classified, or condensed in any way whatsoever.

Secondary data are primary data which have been worked over; that is, subjected to one or more operations the extent and character of which are not obvious. As a result of these operations, certain details of the original observations are lost. Thus, in table 2, of the 1,244 laborers who are tabulated as receiving 50¢ and under 60¢, all record is lost of the distribution within the class interval; that is, all record is lost of how many received 55¢ per hour, how many 56¢ per hour, and so on. There is no evidence left of (1) what adjustments had to be made in the data as originally collected, (2) the purpose for which they were collected, or (3) the manner in which they were edited.

Data which are primary in the hands of one individual may become secondary for others. Thus, the Decennial Census data in the hands of the U. S. Census Bureau are primary data. The published reports of the Bureau become a source of secondary data for other investigators.

**7. Primary data.**—An intelligent use of secondary data demands an appreciation of the difficulties under which primary data are obtained. The statements made here on this subject are not intended to be complete or exhaustive. For more complete statements we refer the reader to other writers.[1]

[1] Crum and Patton, "Economic Statistics," Chapter IV, A. W. Shaw Co.
Jerome, Harry, "Statistical Methods," Chapter XVI, Harper's.
Secrist, Horace, "An Introduction to Statistical Methods," Chapter III, revised edition, Macmillan Co.

Primary data are obtained to test a specific theory which has been formed with respect to a particular problem. Before collecting any primary data, the observer should first examine all available secondary data bearing on the problem and from this examination determine what additional data are needed. Having determined this, and the tabular form in which they should be presented, the investigator has next to determine where and from whom the data can be obtained and the method to be used in collecting the data.

Shall the collection of data be complete or shall a sample only be obtained? If a sample, how large shall the sample be, and what precautions must be taken to insure that the sample be representative? The question of time and expense is involved. If the data are *registration data,* one must determine the place of registration and obtain access to the register. Is the registration complete, accurate, and reliable? Deaths, births, marriages, and divorces are matters of registration. Income tax returns are of permanent record. The returns to the Department of Agriculture of its crop reporting service are a matter of record. Banks make periodic statements of conditions to the Comptroller of the Currency. There are permanent records of bank failures and bank clearings.

If the data required are not a matter of registration, they must be obtained otherwise. In general, four different plans may be used in collecting primary data. They are: (1) by personal inquiry, in the nature of a house to house canvass; (2) by enumerators or special agents; (3) by a mailed questionnaire; and (4) by obtaining estimates from correspondents.

**8. Personal inquiry.**—This method can be used to advantage when it is desired to study a few cases intensively and in a short period of time. It may be used also in getting general information and surveying the lay of the land before mapping out a fuller and more extensive campaign. Information secured by personal inquiry should be comparatively easy to obtain. The interviewer who has a personal interest in the problem to be studied and who is concerned with the accuracy of the results should be able to obtain personal information, opinions, theories, suggestions, and comments that may not be secured in any other way. The data secured by a properly qualified interviewer may serve to supplement and to correct information already on hand. However, interviews are often superficial and general. Only a small number of cases can be studied, and there is danger that the principles set forth as a result of such inquiry may not be truly representative of the entire group from which the samples are taken. There is also grave danger that results may be biased and that the prejudices and desires of the investigator may, without his being aware of it, affect his conclusions.

**9. Enumerators or special agents.**—If the number to be interviewed is large or the investigation extensive, then one has to decide as to whether enumerators shall be selected or a questionnaire mailed. In either case forms must be prepared. The units of measurement must be clearly stated and defined. For example, are days of employment given in number of days per week or number of days per month? What constitutes a day's work? Is it 8 or 10 hours' work? Honest and capable enumerators must be chosen. They must be given complete instructions as to the purpose of the survey and the purport of each question to which they are to obtain an answer. Thousands of trained enumerators are used by the United States Government in taking the census of population. The census of a city or of the entire nation may be taken by this method in about two weeks. The method is satisfactory for extensive studies, but it is entirely too expensive for most kinds of intensive inquiries.

**10. The questionnaire.**—The questionnaire method is one of the most widely used methods of collecting statistical data. Its success lies in the fact that an extensive inquiry can be made and much material gathered with less expense than by any other method. Questionnaires are usually sent by mail, but may be taken to the informant in person and left to be filled out and sent in at a later date. They may take the form of a letter, a double postcard, or may be printed in advertisements.

Certain rules should be followed in preparing questionnaires. The questions to which answers are to be obtained should be adequate, but few in number, and as brief as is consistent with clearness. Questions should be such as are capable of statistical study and so framed that they can be answered by *yes, no,* or a number. Leading questions should not be asked, because they are likely to arouse suspicion or animosity. Indefinite questions may be used occasionally to draw out opinions. Check questions should be used. Thus, if one's age is asked, the date of birth should also be asked. In obtaining nationality, the place of birth should also be obtained. Clearness is perhaps the most important guiding principle to be observed. The choice of words should receive much attention. The schedule may not be returned or the answers may not be correct if the questions are misunderstood or if the units of measurement are not clearly set forth.

In framing questionnaires it is common practice to "cloak" the real objective. When information of a confidential nature is desired and it is believed that the informant will not furnish complete and accurate data in answering a single question, a series of questions may be asked which seemingly have no relation to each other. For example, let us assume that one desires to know the total amounts appropriated for advertising by leading department stores in Kansas. A series of three questions may be inserted among a larger number and a questionnaire mailed. One question may have to do with the percentage of appropriation distributed among certain enumerated media; another with the number of newspapers used by the store for advertising purposes, and a third with the average amount spent in each newspaper carrying the store's advertising. With this information at hand, the amount of the entire appropriation can, of course, be computed. There are dangers in this method which should be apparent to the reader and which should serve to discourage its use in the securing of confidential data.

The questionnaire method has a serious defect in that it depends to a large extent upon disinterested persons for information. Only a small number of questionnaires are returned, and of this number many are incomplete and full of errors. Much information that is received is unsatisfactory because of hasty or perfunctory replies. Other objections to the questionnaire method are: (1) that the majority of people hesitate to commit themselves in writing; (2) that selfish motives sometimes influence replies; (3) that a fair sample of data is not secured even though the original list of informants may have been made very carefully; and (4) that complete information is not secured, because of lack of the stimulus of personal contact. Only by careful checking and follow-up can one hope to get satisfactory results from the use of this method.

**11. Estimates from correspondents.**—This method is used in conducting extensive investigations and differs from the questionnaire method in that the informant is presumed to have no accurate knowledge of the questions asked. Answers are based upon facts which are *estimated* and not *counted.* The method is used by the United States Government in obtaining crop reports and is favored because of its simplicity and inexpensiveness.

**12. Editing.**—The returns must be edited for completeness, consistency, uniformity, and accuracy. A return is incomplete if some of the questions are unanswered. In this case one must either interview the individual again or decide whether the incomplete return can be used as it stands or must be thrown out. If the reported age and the age as computed from date of birth do not agree, the returns are inconsistent, and either the inconsistency must be removed or this return rejected in whole or in part. Uniformity must be obtained with reference to the units employed. Lack of uniformity can usually be detected and remedied by the investigator without further investigation. For example, if daily wage is reported where monthly wage is asked for, the investigator can make the needed adjustment. Other inaccuracies are difficult to detect and, when detected, may cause much difficulty in correcting. However, these inaccuracies must be corrected or the return not tabulated. The trustworthiness of the whole set of data and the ultimate usefulness of the data depend upon the ability and honesty of the editor in detecting and remedying inaccuracies in the returns.

**13. Secondary data.**—A statistical investigator should become familiar with the sources of secondary data. The secondary data are examined first in making a statistical investigation. We will give here a brief outline of some of the chief sources of secondary data. These may be classified as:

I. United States Government publications.
II. Commercial and trade papers.
III. Daily papers.
IV. Private statistical service.
V. State publications.
VI. Research agencies.
VII. Trade associations.
VIII. Yearbooks.

We list a few of the more important sources under each of these main divisions:

I. United States Government publications:

A. United States Department of Commerce:
1. Bureau of the Census:
a. Decennial Census. Fourteenth census of the United States, 1920, XI volumes.
Population:
Vol. I.—Number and Distribution of Inhabitants.
Vol. II.—General Report and Analytical Tables.
Vol. III.—Composition and Characteristics of the Population, by States.
Vol. IV.—Occupations.
Agriculture:
Vol. V.—General Report and Analytical Tables.
Vol. VI.—Reports for States, with Statistics for Counties.
Vol. VII.—Irrigation and Drainage.
Manufactures:
Vol. VIII.—General Report and Analytical Tables.
Vol. IX.—Reports for States, with Statistics for Principal Cities.
Vol. X.—Reports for Selected Industries.
Mining:
Vol. XI.—Mines and Quarries.

b. Annual Publications:
1. Mortality Statistics.
2. Birth Statistics for the Birth Registration Area.

3 Financial Statistics of Cities.
4. Financial Statistics of States.
c. Special Reports:

2. Bureau of Foreign and Domestic Commerce:
a. Statistical Abstract of the United States.
b. Commerce Yearbook.
c. Commerce Reports (Weekly).
d. Monthly Summary of Foreign Commerce of the United States.

B. United States Department of Agriculture:
1. Yearbook (Annual).
2. Weather, Crops, and Markets (Weekly).

C. United States Department of Labor:
1. Bureau of Immigration.
a. Annual Report of the Commissioner General.
2. Bureau of Labor Statistics:
a. Monthly Labor Review.
b. Retail Prices and Cost of Living Series.
c. Wholesale Price Series.
d. Wages and Hours of Labor Series.

D. United States Department of the Interior:
1. Bureau of Mines.
2. Geological Survey.
Mineral Resources of the United States (Annual).

E. Federal Reserve Board:
1. Annual Report.
2. Federal Reserve Bulletin (Monthly).

F. Interstate Commerce Commission:
Statistics of Railways in the United States (Annual).
G. Treasury Department.

II. Commercial and trade papers:
A. The Annalist.
B. Bradstreet's.
C. Dun's Review.
D. Commercial and Financial Chronicle.
E. The Economist (London).
F. The Statist (London).
G. L'Economiste Français (Paris).
H. Trade Journals:
1. Coal Age.
2. Iron Age.
3. Railway Age.

III. Daily papers:
A. Wall Street Journal.
B. Journal of Commerce (New York).
C. General Daily Newspapers (Financial Columns).

IV. Private statistical service:
- A. Babson's Statistical Organization, Babson Park, Mass.
- B. Brookmire Economic Service, New York.
- C. Harvard Economic Service, Cambridge, Mass.
- D. Moody's Investors' Service, New York, N. Y.

V. State publications:
- A. Illinois Department of Labor, Springfield.
- B. Massachusetts Department of Labor and Industries, Boston.
- C. New York State Industrial Commission, Albany.
- D. Wisconsin Industrial Commission, Madison.

VI. Research agencies:[1]
- A. National Bureau of Economic Research, 474 West Twenty-fourth Street, New York, N. Y.
- B. Food Research Institute, Stanford University, Palo Alto, California.

VII. Trade associations.[2]

VIII. Yearbooks:
- A. The American Yearbook, Appleton, New York.
- B. The Statesman's Yearbook, Macmillan, New York.
- C. The World Almanac and Encyclopaedia, The Press Publishing Co., The New York World, New York.

**14. Tests to be applied to secondary data.**—Before making use of secondary data one should determine the reliability of the data. A number of factors need to be considered in this connection.

One should determine through what agency the data are collected. Does this agency have access to all of the data bearing on the question? What are the standards of excellence of the collecting agency? Does the agency have any interest in the results of the investigation?

Determine whether the data represent a complete enumeration or a random sample. If the data represent a sample, is the sample large enough to be characteristic of the whole group? Thus, the proportion of negroes to whites in New Orleans would not characterize conditions in the cities of the United States as a whole. The ratio of millionaires in Colorado Springs to the total population of that city would

[1] See "Market Research Agencies," a 156-page pamphlet, 1927 edition, by the United States Department of Commerce. 737 research agencies with the addresses are listed.

[2] E. H. Naylor, "Trade Associations," Ronald Press, 1921, pp. 348–379. Quite a complete list with addresses.

not be typical for cities of a population of 40,000 in the United States. If the data represent a sample, is the sample taken at a time which is truly representative and from a group which is truly representative? From whatever source it might enter, it must be determined whether or not bias is present.

Difficulties arise in connection with the units employed. The definition of a unit may change from time to time and from place to place. This difficulty will be overcome if you try to define for yourself what is meant by a private dwelling, an improved farm, or a paved road. In stating monthly wages one report may use four weeks as the unit, another report may use thirty days. It is difficult to compare tax rates, as the rate in one city may be based on an assessed valuation of 50 per cent of actual value while in another city the rate may be based on 75 per cent of actual value. This rate of assessed to actual value may change in any given city from time to time. In computing an index number of retail prices, the United States Bureau of Labor Statistics included after 1907 fifteen commodities not included before. Thus, index numbers before and after this time are not, strictly speaking, comparable.

In statistics of railroads, monthly data on revenues and expenses of steam railroads after October, 1910, included switching and terminal companies, but not before. Thus, these figures cannot well be compared.

One must determine the *accuracy* of the data. If age is given, is it correctly reported? This can at times be determined from internal evidence and methods devised for smoothing the tabulated returns. Wages may be falsely reported through aroused suspicions as to the use to which the data will be put. In medical statistics inaccuracies arise from a lack of uniformity in nomenclature. Quoting from Raymond Pearl in "Medical Biometry and Statistics," Chapter III, p. 58:

> Some physicians all the time, and all physicians some of the time, will use their own terminology instead of that of the International

Classification in reporting the cause of death on the original death certificate.

There is no adequate reporting service at the present time on gold production in the United States. Any figures given are mere rough estimates and are therefore inaccurate. The same can be said with respect to the amount of standing timber.

It should now be clear that one must exercise considerable care in selecting secondary data for use in any statistical investigation.

## Chapter III

## SAMPLING

**15. Introduction.**—One of the most frequent problems in statistical method is to determine, from a properly chosen sample, the composition of the whole[1] from which the sample was taken. In many cases it is impossible or impractical to make a complete count. For example, it is impossible to make a complete count of all mice for the determination of any characteristic whatever. It is impractical to make a study of the living conditions of every steel worker in Gary, Indiana. Inasmuch as we are to determine from the sample the composition of the whole, we must endeavor to take the sample in such a way that it will contain the characteristics of the whole in the same proportion as the whole. We are concerned, then, with the following questions with respect to a sample:

1. Is the sample a good one?

2. Is the sample large enough?

3. What reliability is to be attached to the measures computed from the sample?

**16. Laws of sampling.**—A complete count would be necessary if it were not for the uniformity observed in nature. Due to this uniformity, we expect measures computed from different samples to approach closely the true measure for the universe from which the samples were chosen.

We expect measures computed from different samples to differ from each other by amounts which are comparatively small.

[1] The aggregate from which the sample is taken is commonly referred to as the "Universe" of discourse, or the "Population" under discussion.

Variations do occur in nature, but these variations are not limitless; they are confined within certain definite limits. The heights of men do differ, but the range of this variation is comparatively small.

There are two interesting and important manifestations of this uniformity in nature. These are:

1. The stability of large numbers.
2. The permanence of small numbers.

1. *Stability of large numbers.*—In the absence of any cause tending to produce a decided change, from time to time or from place to place, the number of happenings of an event out of a fixed large number of possible happenings remains about constant.

This is the basis of all modern insurance. From 100,000 people living at age 10, experience has shown that 74,985 will live to celebrate their 44th birthday and that, of these, 812 will die before reaching the age of 45. These numbers are used to compute the natural premium on a term insurance policy issued to one of age 10 which expires at age 45.

As an additional illustration we give here the number of births and deaths per 1,000 of population in the birth registration area, as given on p. 68 of the Statistical Abstract of the United States for 1924.

**Table 5.—Births and Deaths per 1,000 Population.**

| *Births per 1,000 Population* | | | | | *Deaths per 1,000 Population* | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1920 | 1921 | 1922 | 1923 | 1924 | 1920 | 1921 | 1922 | 1923 | 1924 |
| 23.7 | 24.3 | 22.5 | 22.4 | 22.5 | 13.1 | 11.7 | 11.9 | 12.4 | 11.9 |

This principle is again exemplified in the number of millions of swine on farms and in the total production in millions of bushels of corn and wheat in the United States from year to year as shown by the following data taken from the Statistical Abstract of the United States for 1924, pp. 587 and 608.

**Table 6.—Production of Swine, Corn, and Wheat.**

| | *1913* | *1915* | *1917* | *1919* | *1921* | *1922* | *1923* | *1924* |
|---|---|---|---|---|---|---|---|---|
| *Swine*............ | 61.1 | 64.6 | 67.5 | 74.6 | 56.1 | 58.3 | 68.4 | 66.1 |
| *Corn*............. | 2,447 | 2,995 | 3,065 | 2,811 | 3,069 | 2,906 | 3,054 | 2,437 |
| *Wheat*............ | 763 | 1,026 | 637 | 968 | 815 | 868 | 797 | 873 |

The years 1915 and 1917 were, during the World War, a time when great effort was being made to increase production; yet the production did not fluctuate much more than it has in the past in times of peace. The number of swine in 1909 was 54,147,000 and in 1911 was 65,620,000. The production of wheat was 700,434,000 bu. in 1909 and 621,338,000 bu. in 1911. The production of corn was 2,531 millions of bu. in 1911 and 3,125 millions of bu. in 1912.

It takes time materially to increase the number of swine. Thus, the greatest increase was not registered until 1919. The census was taken in 1910 and not again until 1920. The intervening figures are estimates.

2. *The permanence of small numbers.*—If the number of happenings of an event out of a fixed large number of possible happenings is small in one instance, then we may expect that in other instances, out of an equally large number of possible happenings, the number of happenings will be small. The variations in this small number will be slight. The number seldom vanishes.

In reference to this principle Bowley[2] says:

> Specialists in all professions, from the doctor who treats only one obscure disease of the ear to the dealer in curiosities, make their livelihood dependent on this permanence of small numbers.

For example, the number suffering from a given disease of the skin is comparatively small. Any city of 35,000 population with its contributing adjacent area will support no more than one skin specialist. Within this district the specialist

[2] A. L. Bowley, "Elements of Statistics," 4th ed., p 286, Scribner's, N. Y., 1921.

will encounter, on the average, one and possibly two cases each year of *lichen planas.*

This principle is amply verified and illustrated by the following data on accidents while operating switches on class I railroads:

| 1919 | 1920 | 1921 | 1922 | 1923 | 1924 |
|---|---|---|---|---|---|
| 3 | 2 | 2 | 1 | 1 | 3 |

*Bureau of Labor Statistics, Bulletin No. 425, p. 44.*

It is remarkable that, from among the thousands of switches operated each year, in no one of these six years were there more than three accidents, and never less than one.

As further illustrations we have the data given in tables 7 and 8.

**Table 7.—Actual Number of Deaths for the Ten Year Period 1910–1920.**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Glanders*.................. | 1 | 1 | 4 | 4 | 4 | 4 | 10 | 8 | 12 | 11 |
| *Rabies*.................... | 41 | 58 | 63 | 66 | 36 | 52 | 65 | 95 | 74 | 83 |

*Bureau of Census, Mortality Statistics, 1920.*

**Table 8.—Illustration of Permanence of Small Numbers.**

| | *1915* | *1920* | *1921* | *1922* | *1923* | *1924* | *1925* |
|---|---|---|---|---|---|---|---|
| COAL MINES | | | | | | | |
| *Shaft fatalities*............ | 40 | 56 | 36 | 41 | 46 | 29 | 34 |
| METAL MINES | | | | | | | |
| *Rock or ore underground loading*................ | 10 | 9 | 2 | 3 | 2 | 8 | |
| *Electricity* ............... | 14 | 18 | 4 | 7 | 12 | 13 | |
| *Run or Fall of Ore*........ | 12 | 7 | 6 | 7 | 6 | 4 | |
| SMELTING PLANTS | | | | | | | |
| *Falls of persons*........... | 3 | 0 | 0 | 5 | 3 | 1 | |
| *Flying or falling objects*.... | 1 | 0 | 0 | 1 | 1 | 2 | |
| QUARRIES | | | | | | | |
| *Drilling and channeling*.... | 1 | 2 | 1 | 1 | 0 | 1 | |

*Statistical Abstract of U. S., 1925, pp. 743–744.*

As shown by table 8, the number of happenings of an event does at times become zero. It is to be noticed that all of the numbers in any one row are of the same order of magnitude.

**17. Conditions of simple sampling.**—What are the conditions under which one may expect a sample to possess, within certain ascertainable limits, the characteristics of the universe from which the sample was chosen?

(a) *Independence.*—The various items of the sample must be independent. Every item in the population under consideration must have the same chance for inclusion in the sample. The selection of one item for the sample must not change the chances for the selection of other items.

The drawings of balls from a bag are not independent. If the bag contained originally one hundred balls, then the chances for selection of the first ball is 1 in 100. After the first ball is chosen only 99 remain and the chances for the selection for the next ball becomes 1 in 99.

The death rate due to typhoid for El Paso Co., Colorado, might be three times the death rate for the United States as a whole. This sample then would not be representative of the United States as a whole. The presence of any infectious or contagious disease in one person increases the possibility of others near him contracting the disease, with the attendant possibility of death.

Deaths due to a single railway accident are not independent events. The chances of death are not the same for those on the train and for those not on the train.

Successive tosses of a coin are independent events. The appearance of heads on the first throw has no influence as to what turns up on the next throw or any other throw.

(b) *Homogeneity.*—The death rates for a group of age 16 cannot be applied to a group of age 70. The death rates for persons aged 70 should be determined from a sample of persons aged 70. The death rates for men aged 70 should be determined from a sample of men aged 70 and not from a sample of women aged 70.

To determine the rate for fire insurance in the business district of Kansas City, it is not necessary to investigate every building. A sample only is taken. This study, however, would give no indication as to what the fire hazards are in the Denver business district. Conditions in the two cities might be radically different. Localities with the same general conditions are given the same rate.

The theory of sampling cannot be applied to death rates for a given disease in successive years during a period in which continued advancements in medical science have brought about a continuous decrease in the death rate. Thus the death rate due to tuberculosis today is not indicative of what the death rate was ten years ago or what it may be in the future.

On this point Yule[3] says:

> There must not be a difference in any essential respect—*i.e.*, in any character that can affect the proportion observed—between the localities from which the observations are drawn, nor, if the observations have been made at different epochs, must any essential change have taken place during the period over which the observations are spread.

**18. How to take a sample.**—Considerable care must be exercised to insure that the sample is representative of the whole from which it is taken.

*Random selection.*—If the extent of the population is known, a representative sample may be obtained in the following manner: Write numbers on slips of paper, one for each individual in the population. Place the slips of paper in a box, shake well, and draw one out; shake well and draw another slip , and so on. By the use of this plan, all the individuals have nearly equal chances of inclusion in the sample. If we are dealing with a population of 1,000, the chances for inclusion on the first draw are 1 in 1,000. If the sample is to contain 100, the chances for inclusion on the last draw are 1 in 900. This is the familiar process of drawing by lot.

*Selection by design.*—The sample must contain individuals from each of the classes which constitute the universe, and

[3] G. U. Yule, "An Introduction to the Theory of Statistics," p. 260, C. Griffin and Co., Ltd., London.

in the same proportion. If a population sample is desired for a given district, some measures must be taken to insure that all age classes are represented in the sample. If the classes for age are 0–10, 10–20, 20–30, 30–40, 40–50, 50–60, 60–70, 70–80, 80–90, 90–100, the sample must contain not less than 10 individuals, one from each age class. If both males and females are present in the district, then at least one male and one female should be represented in each age class. This would bring our sample to not less than 20 individuals. If white, black, and brown races are present in the district, our sample would be increased to no less than 60 individuals.

If a sample is to be taken over a given geographical area, a mesh can be laid down over the area and one sample taken from each mesh.

In sampling ore, a small quantity is taken at regular intervals. These samples are thoroughly mixed and a small portion of the mixture assayed by a chemist.

If two samples differ largely, then either the sample is not large enough or there is considerable heterogeneity in the original material and both samples are probably biased by not being representative of the whole. Thus, if salmon from the Columbia river differ in one or more characteristics from salmon from the Fraser river, then a study of Columbia river salmon would not be representative of salmon as a whole.

If the investigation is one concerned with the living conditions in a factory town, members from each wage group must be included, in the *proportion* in which they exist to the entire population. The different racial elements must be included. The investigation must not be confined to those who keep a record of incomes and expenditures.

If the items are named, then a sample can be obtained by arranging the names alphabetically and taking every tenth individual.

**19. Reliability of a sample.**—As stated above, if two samples differ largely, then either the sample is not large enough or there is considerable heterogeneity in the original material. Let us assume that the material is homogeneous.

A sample is large enough whenever several samples of the same size present substantially the same characteristics.

Take[4] a sample of 400 ears of corn from a crib. Compute the arithmetic average of their lengths. We use this computed length to represent the arithmetic average length of all of the ears in the crib. The true arithmetic average length of all of the ears in the crib may differ from this computed average. There is a number, which we shall designate by the symbol $E_m$, called the *probable error in the arithmetic average.* This is a number such that the chances are even that the computed arithmetic average lies between $\overline{X} - E_m$ and $\overline{X} + E_m$, where $\overline{X}$ denotes the unknown true arithmetic average.

This means that if a large number of persons take a sample of ears and each computes an average length, then about one half of these averages will be within the limits set and one half will be without.

In treatises on probability it is shown that

$$E_m = \frac{0.6745\sigma}{\sqrt{n}}$$

where $n$ is the number of items in the sample.

This formula shows that the probable error in an arithmetic average obtained from a sample varies inversely as the square root of the number of items in the sample. In other words, in order to double the precision of the computed arithmetic average, it is necessary to include four times as many items in the sample; to treble the precision, it is necessary to include nine times as many items in the sample. In order to double the precision of an arithmetic average computed from 400 ears, it becomes necessary to include 1,600 ears in the sample.

One of the fundamental problems of statistics is that of determining the limits to the fluctuations in the various constants computed from a sample.[5]

[4] Adapted from Kenyon-Lovitt, "Mathematics for Agriculture and General Science," p. 301, The Macmillan Company, New York.

[5] For a list of the probable errors in various statistical constants, consult H. L. Rietz, "Handbook of Mathematical Statistics," p. 77. Houghton Mifflin Co., Boston.

**20. Units of measurement.**—In taking a sample, one must carefully select and define the units in which the measurements are to be taken. If one is measuring the volume of business, should one choose the dollar or the ton as a proper unit of measure? In measuring passenger service, should the unit be the passenger or a passenger-mile? In measuring freight service, should the data be given in terms of tons handled, ton-miles, or dollars of revenue?

If one is recording industrial accidents, what shall be recorded? Twenty states record all accidents. Nine states record only those accidents which give rise to one or more days' disability. One state records only those accidents which give rise to two or more days' disability. Others require reports only on accidents involving more than one week's disability.

A question of fundamental importance in many cases is whether the unit of measurement should be one of value or one of quantity. For example, which is more significant, the value of the wheat consumed, or the quantity of wheat consumed?

### Exercises.

1. A factory has 5 different wage groups. In each group there are males and females of three different races, namely, Scandinavian, Italian, and Russian. What is the minimum number of the following classes which should be included in a sample in a study of wages: Men, women, men and women, Italians, Italian men, Italian women, each wage group?

2. If $0.6745\sigma$ is 1, how many items must be included in a sample in order that the probable error of the arithmetic average shall not exceed (a) 0.1; (b) 0.05; (c) 0.01; (d) 0.2; (e) $\frac{1}{12}$; (f) $\frac{1}{24}$; (g) $\frac{1}{6}$; (h) $\frac{1}{3}$; (i) $\frac{2}{3}$; (j) 0.3; (k) 0.04?

3. Bring in additional illustrations of: (a) the stability of large numbers, and (b) the permanence of small numbers.

4. A sample of 100 ears of corn has an arithmetic average length of ear of 10 inches. The probable error is 0.5 inch. Another sample of 100 ears from the same field has an average length of 9 inches. Is the sample large enough? *Ans.* No.

## Chapter IV

# TABULATION

**21. Introduction.**—The first step in the presentation of collected data is to put them into tabular form. The process of arranging the data in an orderly manner into rows and columns capable of being read in two directions is called *tabulation.* The title at the top of each column is called a *caption.* The title at the left of each row is called a *stub.*

An orderly arrangement of data into groups having a common characteristic is a first step in any scientific study of the data. Having classified the data into groups, we are in a position to make comparisons or contrasts, and to draw inferences. Tabulations are made with the hope that this orderly arrangement may suggest relationships or at least help us to detect relationships which are not apparent. Tabulation is the first step towards a visualization of the relations existing between different parts of the data.

Tabulation is a summarization process. It requires that the characteristics to be presented must be carefully determined and a satisfactory scheme of grouping for each characteristic be fixed. For example, if one of the tabulated items is wages, to the nearest cent, the distribution may be so wide as to preclude the possibility of separate entries for each distinct wage amount received. A satisfactory scheme of grouping having been determined, each item can be assigned to its proper group. Each group must be specified exactly. There must be no border-line cases. It is desirable that all groups be of the same width. The number of groups should be such as to avoid excessive summarization.

**22. Order of arrangement.**—Groups of items are commonly arranged according to size with the size increasing from left to right and from top to bottom. On occasion the magnitude increases from bottom to top to agree with the order for a plotted diagram. We illustrate with the following data from the Statistical Abstract of the United States, 1924, p. 564:

**Table 9.—Number of Farms by Size, 1920.**

| *Size* | *Number* |
|---|---|
| Under 10 acres | 288,772 |
| 10 to 19 acres | 507,763 |
| 20 to 49 acres | 1,503,732 |
| 50 to 99 acres | 1,474,745 |
| 100 to 499 acres | 2,456,107 |
| 500 to 999 acres | 149,819 |
| 1,000 acres and over | 67,405 |
| Total | 6,448,343 |

In table 10 the magnitude increases from left to right.

**Table 10.—Frequencies of Numbers of Petals for Ranunculus Bulbosus.**

| *Number of Petals* | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|
| *Frequency* | 133 | 55 | 23 | 7 | 2 | 2 |

*H. De Vries, Ber. Dtsch. Bot. Ges., Bd. XII, 1894.*

If one group of items refers to time, then time increases from left to right and from top to bottom. The sequence of time must not be broken. Table 11 illustrates this.

**Table 11.—Average United States Farm Price for Eggs, in Cents per Dozen.**

| *Year* | *Price* |
|---|---|
| 1921 | 33 |
| 1922 | 28 |
| 1923 | 30 |
| 1924 | 25 |

| Month 1924 | *Jan.* | *Feb.* | *March* | *April* | *May* | *June* | *July* | *Aug.* | *Sept.* | *Oct.* | *Nov.* | *Dec.* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Price | 35 | 34 | 20 | 19 | 20 | 21 | 23 | 26 | 32 | 38 | 46 | 50 |

*Statistical Abstract of U. S., 1924, p. 602.*

If the data tabulated refer to spatial items, the practice is to tabulate the items in the order of their occurrence in space, naming consecutively, in so far as it is possible, contiguous items. Thus, for data concerning the various states of the United States, the states would, in general, not be listed alphabetically but from East to West, following a well established order used by the United States Census Bureau. This order might very well be altered, however, if one were considering the distribution of Mexicans within the United States. Table 12 gives the order of the states as used by the United States Census Bureau.

**Table 12.—Corn Yield per Acre by States, 1923.**

| State | Yield | State | Yield | State | Yield |
|---|---|---|---|---|---|
| *New England* | 40.6 | *East North Central* | 38.1 | *West South Central* | 15.7 |
| Maine | 38.0 | Ohio | 41.0 | Louisiana | 15.4 |
| New Hampshire | 42.0 | Indiana | 38.5 | Texas | 18.5 |
| Vermont | 39.0 | Illinois | 37.5 | Oklahoma | 11.5 |
| Massachusetts | 43.0 | Michigan | 34.5 | Arkansas | 15.5 |
| Rhode Island | 38.0 | Wisconsin | 37.0 | | |
| Connecticut | 41.0 | | | | |
| *Middle Atlantic* | 37.7 | *West North Central* | 33.4 | *Mountain* | 25.1 |
| New York | 32.4 | Minnesota | 36.0 | Montana | 26.0 |
| New Jersey | 40.0 | Iowa | 40.5 | Wyoming | 27.0 |
| Pennsylvania | 40.0 | Missouri | 30.0 | Colorado | 25.0 |
| | | North Dakota | 33.5 | New Mexico | 16.4 |
| | | South Dakota | 34.5 | Arizona | 30.0 |
| | | Nebraska | 33.0 | Utah | 24.9 |
| | | Kansas | 21.7 | Nevada | 23.3 |
| | | | | Idaho | 42.0 |
| *South Atlantic* | 20.2 | *East South Central* | 20.8 | *Pacific* | 35.5 |
| Delaware | 33.1 | Kentucky | 28.5 | Washington | 37.0 |
| Maryland | 39.3 | Tennessee | 24.5 | Oregon | 35.0 |
| Virginia | 29.0 | Alabama | 14.0 | California | 35.0 |
| West Virginia | 34.0 | Mississippi | 14.5 | | |
| North Carolina | 22.5 | | | | |
| South Carolina | 16.5 | | | | |
| Georgia | 12.2 | | | | |
| Florida | 12.5 | | | | |

*Statistical Abstract of U. S., 1924, p. 624.*

In general, in any tabulated data, the order of the items should conform as nearly as is possible to the relationship, casual or otherwise, existing between the items.

Frequency tables are constructed by listing groups in an increasing order with respect to the measured characteristic and by placing opposite each group the frequency of occurrence of the characteristic with the stated measurement. Thus table 10, which gives the number of petals for *ranunculus bulbosus*, is a frequency table.

**23. Class interval.**—In choosing a class interval for the groups, the following considerations should be kept in mind.

(1) The size of the class interval should be such that the characteristic features of the distribution are displayed.

(2) Tabulation is a summarizing process. The class intervals taken must not be so large that a considerable error would be involved in assuming that the arithmetic average[1] of the items within the interval is the mid-point of the interval. We lay down the general rule that the number of classes should never be less than six and usually should not be less than fifteen nor more than thirty.

(3) The interval should not be so small as to give, within the significant parts of the range, class intervals with a zero frequency. Remember that summarization is desired. Do not make the interval so small as to lose the advantages due to summarization. Usually twenty-five to thirty classes are enough. For a frequency distribution, the interval should be such as to secure some regularity in the frequencies.

(4) The class interval should be of a *uniform* width. This will aid in making a graph representing the series. If the class intervals are uniform, this interval may be taken as a unit of measure in computing an average by the step-deviation process.

(5) A comparison of different distributions is facilitated if the same class interval is used for all.

(6) In general, there should be no class intervals without definite limits. There should be no intervals of the type "all over" or "all under."

**24. Class limits.**—It is convenient in computation and tabulation to have integers for class limits. For computational purposes it is desirable to have an integer as the

[1] For a discussion of arithmetic average, see chapter VII.

mid-point of an interval. Both of these things can be obtained by making the class interval an odd number of units. Five units is a convenient interval width. When integers are not advisable for class limits, try to have the limits represented by simple fractions.

The class limits must be unambiguous. Each group must be specified exactly. There must be no border line cases.

In the case of a discrete variable, the class limits are ordinarily evident. Thus, if coal mines are to be classified with respect to the number of pick and shovel employees, and the classes are 1–100, 101–200, 201–300, etc., it is clear that the lower limits of the classes are 1, 101, 201, etc., while 100, 200, 300, etc., are the upper limits of the classes.

In the case of a continuous variable, the class limits and class boundaries may not be so clear. Suppose the classes for weights of men in pounds were as follows:

120—130
131—141
142—152
153—163
164—174
175—185
186—196
197—207

It is clear that the lower limits are 120, 131, 142, etc., and the upper limits are 130, 141, 152, etc. The class boundaries are 130.5, 141.5, 152.5, etc. A man who weighed 152.4 pounds would be put in the third class. A man who weighed 152.6 pounds would be put in the fourth class. Of those men who weighed 152.5 pounds, half would be put in the third class and half in the fourth class. Some prefer to place all of these border line cases in the higher group.

The weight classes given above might be given in any one of the following forms:

|  |  | *Mid-point* |
|---|---|---|
| *(1)* | *(2)* | *(3)* |
| 119.5 and under 130.5 | 119.5—130.5 | 125 |
| 130.5 and under 141.5 | 130.5—141.5 | 136 |
| 141.5 and under 152.5 | 141.5—152.5 | 147 |

Form (1) is clear in meaning but cumbersome. The class limits are not indicated. The class boundaries are 119.5, 130.5, 141.5, etc., as before. The class limits can be found when the accuracy of the measurements is given. If the accuracy is that of the nearest half pound, then the upper limit of the first class is 130 lbs., and the lower limit of the second class is 131 lbs.

Form (2), when used, is usually meant to be read as form (1). For example, 119.5–130.5 is read 119.5 and under 130.5.

When form (3) is used, the class boundaries can be computed. For example:

$$\frac{125 + 136}{2} = 130.5$$

That is, the class boundary is halfway between the midpoints. For the lower class boundary of the first class we have:

$$125 - \frac{136 - 125}{2} = 125 - 5.5 = 119.5$$

The following tables illustrate some of the types mentioned above:

**Table 13.—Petroleum Refining, 1919.**

| *Value of Product* | *Number of Establishments* |
|---|---|
| Less than $5,000 | 4 |
| 5,000 to 20,000 | 6 |
| 20,000 to 100,000 | 21 |
| 100,000 to 500,000 | 65 |
| 500,000 to 1,000,000 | 56 |
| 1,000,000 and over | 165 |

*Fourteenth Census of U. S., 1920, Vol. X, p. 759.*

In this table the class boundaries are clearly stated. The class intervals are not uniform. The lower limit of the first class cannot be less than zero. Actually the lower limit of the first class must be considerably above zero. No indication is given of the location of this lower limit. The last class is left open on the upper end.

For table 14 the class interval is of uniform width except for the first,which is open on the lower end with no indication as to the minimum wage. The class boundaries are clearly given. One interval, namely, 1.50 and under 1.60, is absent. The laborers who received more than $1.00 per hour are so few that the percentage designation becomes unsatisfactory. The frequency for the interval 1.20 and under 1.30 is entirely absent. We cannot be sure that the frequency is zero.

**Table 14.—Per Cent of Laborers Inside Mines with Stated Earnings per Hour, United States, 1924.**

| *Earnings per Hour* | *Laborers inside Mine, %* |
|---|---|
| Under $.30 | 2 |
| 30 and under .40 | 8 |
| 40 " " .50 | 19 |
| 50 " " .60 | 19 |
| 60 " " .70 | 6 |
| 70 " " .80 | 5 |
| 80 " " .90 | 11 |
| 90 " " 1.00 | 29 |
| 1.00 " " 1.10 | (1)[1] |
| 1.10 " " 1.20 | (1) |
| 1.20 " " 1.30 | |
| 1.30 " " 1.40 | (1) |
| 1.40 " " 1.50 | (1) |
| 1.60 " " 1.70 | (1) |

[1] (1) Less than one half of 1 %.

*U. S. Bureau of Labor Statistics, Bulletin No. 416, p. 15.*

**Table 15.—Population of the United States by Ages for 1910.**

| Age, Yrs. | Total, *1910* | Age, Yrs. | Total, *1910* |
|---|---|---|---|
| 17 | 1,786,240 | 30 | 1,854,608 |
| 18 | 1,928,366 | 31 | 1,139,000 |
| 19 | 1,763,061 | 32 | 1,431,468 |
| | | 33 | 1,264,247 |
| 20 | 1,854,622 | 34 | 1,282,862 |
| 21 | 1,789,404 | | |
| 22 | 1,835,060 | 35 | 1,528,717 |
| 23 | 1,791,996 | 36 | 1,255,604 |
| 24 | 1,785,902 | 37 | 1,137,317 |
| | | 38 | 1,361,874 |
| 25 | 1,812,275 | 39 | 1,112,588 |
| 26 | 1,688,385 | | |
| 27 | 1,555,451 | | |
| 28 | 1,729,763 | 40 | 1,520,685 |
| 29 | 1,394,129 | 41 | 833,642 |

*Thirteenth Census of the U. S., Vol. I.*

Table 15 is a part of the table which gives the population by ages for 1910. From this summary table it is obvious that at certain ages the age is not reported correctly. The numbers reported as of ages 20, 30, and 40 are too large. There is also an accumulation at ages 25 and 35. Another curious but obvious accumulation is at the ages 18, 28, and 38.

**25. Orders of tabulation.**—A table is said to be of the first order when but one set of numbers is tabulated against the given classes. This is illustrated by the following brief table, No. 16, of the total population of the United States by census years:

**Table 16.—Population of the United States by Census Years.**

| *Census Year* | *Total Population* |
|---|---|
| 1920 | 105,710,620 |
| 1910 | 91,972,266 |
| 1900 | 75,994,575 |

*Fourteenth Census of U. S., 1920, Vol. II, p. 29.*

A table is said to be of the second order whenever the stub or caption contains two coördinate parts. This is illustrated by the following brief table 17:

**Table 17.—Total White and Negro Population of the United States by Census Years.**

| Census Year | Total Population | |
|---|---|---|
| | *White* | *Negro* |
| 1920 | 94,820,915 | 10,463,131 |
| 1910 | 81,731,957 | 9,827,763 |
| 1900 | 66,809,196 | 8,833,994 |

*Fourteenth Census of U. S., 1920, Vol. II, p. 29.*

A table is said to be of the third order whenever each coördinate part under a stub or caption is again divided. We illustrate by the following table 18:

**Table 18.**—Population of the United States by Sex and Color for the Given Census Years.

| Census Year | Total Population | | | |
|---|---|---|---|---|
| | white | | negro | |
| | *Male* | *Female* | *Male* | *Female* |
| 1920 | 48,430,655 | 46,390,260 | 5,209,436 | 5,253,695 |
| 1910 | 42,178,245 | 39,553,712 | 4,885,881 | 4,941,882 |
| 1900 | 34,201,735 | 32,607,461 | 4,386,547 | 4,447,447 |

*Fourteenth Census of the U. S., 1920, Vol. II, p. 107.*

Further subdivisions of stubs and captions will yield tables of higher order. However, it is seldom advisable to proceed in the subdivision to a table of higher order than the third.

**26. Labels.**—The source of the material in a table should always be given. The units in which the measurements are taken should be specified. The smallest division of the measuring instrument taken into account should be stated. Class intervals should be clearly shown.

All titles should be simple, concise, unambiguous, and should state the essential facts of the table. Titles should be self-explanatory. They should answer the queries as to where, when, what. Captions should be more prominent than sub-captions. They should occupy more space and be in heavier type. This remark applies also to stubs and their subdivisions.

A double ruling is customary at the top of a table, a single or a double ruling at the bottom. Captions should be separated from sub-captions by a single ruling. The sides of a table are left open. Totals are sometimes set off by a double ruling.

## Chapter V

## GRAPHIC REPRESENTATION—BAR CHARTS AND MAPS

**27. Introduction.**—The advertising pages of magazines are filled with pictures, charts, and diagrams designed to arrest the attention of the reader and to convey quickly, and with a minimum of effort on the part of the reader, the essential ideas which it is desired to convey. The statistician constructs diagrams and maps to arrest the interest of the public, and to convey the essential features of the data from which the diagram was derived. These diagrams help to clear up ideas and fix facts. The mind has difficulty in holding a mass of numerical data and grasping their significance. A diagram is a summarization of the data. When a graph is properly made, the essential features of the data are presented, and the details submerged.

**28. Pie charts.**—Pie charts, or circular diagrams, are excellent for publicity purposes. These charts are used when one desires to show relative percentages. They are therefore frequently used in the case of financial data. The whole circle is used to represent the dollar. Charts of this kind are used to represent the distribution of costs in a plant, or the parts of a financial budget.

The following data on production of cotton, when expressed in percentages of the total, are well adapted to representation by a circular diagram. The representation is given in fig. 1.

The chief advantages of the circular chart are its simplicity and its power of arresting the attention of the individual.

The circular chart has some disadvantages. The eye does not readily compare the lengths of the various arcs, the areas of the different sectors, or the angles at the center. When several circular charts are shown together they cannot

**Table 19.—Cotton: World Production, in Bales, 1924.**

| | *Bales* | *% of Total* |
|---|---|---|
| U. S. | 13,627,936 | 58.5 |
| India | 5,069,000 | 21.3 |
| Egypt | 1,507,000 | 6.4 |
| China | 2,179,000 | 9.3 |
| Brazil | 605,000 | 2.6 |
| All others | 297,064 | 1.9 |
| TOTAL | 23,285,000 | 100.0 |

*Yearbook, Dep't of Agriculture, 1925, p. 560.*

be readily compared. The labels at times have to be read at awkward angles. To avoid reading at an angle and also to avoid optical illusions, the labels and percentages should be placed outside the circle.

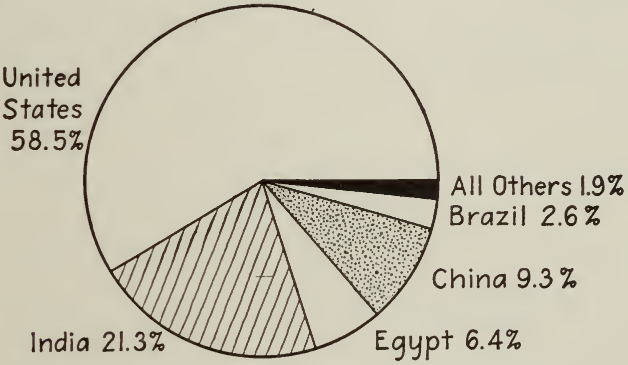


**Fig. 1.—Percentage Distribution of World Production of Cotton in 1924.**

### Exercises.

Make pie charts for the following data (*Source: Statistical Abstract of the U. S., 1924*):

1. *p. 565.* (a) Per cent distribution of tenure of all farm lands.

| | *1920* | *1900* |
|---|---|---|
| Owners | 66.6 | 66.3 |
| Managers | 5.7 | 10.4 |
| Tenants | 27.7 | 23.3 |

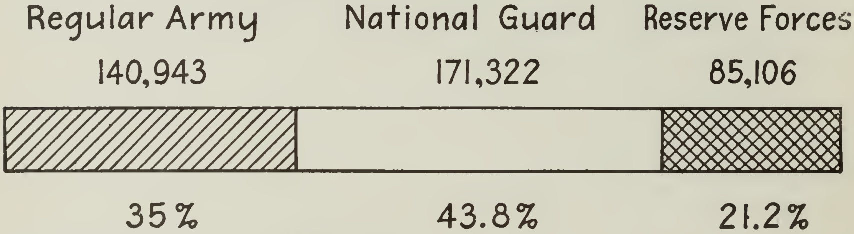
(b) Per cent distribution of all farm lands by color of farmer.

| | *1920* | *1900* |
|---|---|---|
| White | 95.3 | 95.0 |
| Colored | 4.7 | 5.0 |

2. *p. 561.* Per cent of value of all farm property represented by:

| | *1920* | *1910* | *1900* |
|---|---|---|---|
| Land | 70.4 | 69.5 | 63.9 |
| Buildings | 17.4 | 15.4 | 17.4 |
| Implements and Machinery | 4.6 | 3.1 | 3.7 |
| Live stock | 10.3 | 12.0 | 15.0 |

**29. Hundred per cent bars.**—This is one of the simplest and in some respects the best form of graph. This graph consists of a single bar of any convenient length to represent the whole, or one hundred per cent. The bar is subdivided to show what per cent each part is of the whole. The bar is of uniform width, usually from one-tenth to one-twentieth of its length, of such dimensions as will enable one to visualize it readily at the distance at which it is to be read. The different parts of the bar should be shaded, colored, or cross-hatched differently. Percentages should be placed below the corresponding part. Absolute values, if given, should be placed above the parts. In no case should a label of any kind be put inside the bar. The total length of the bar is not significant. Use the hundred per cent bar only where the data can be added to form a coherent whole.



*Statistical Abstract of the U. S., 1924, p. 129.*

Fig. 2.—Army of the United States: Strength of Component Parts, 1924.

These remarks are illustrated by the graph in fig. 2.

**Exercises.**

Construct one hundred per cent charts for the following data (*source: Statistical Abstract of the U. S., 1924*):

1. *p.* 424. Imports entered for consumption and duties collected.

| *Year* | *Per Cent Free* | *Per Cent Dutiable* |
|---|---|---|
| 1821 | 3.96 | 96.04 |
| 1881 | 31.09 | 68.91 |
| 1924 | 59.25 | 40.75 |

2. *p.* 425. Percentage distribution of domestic exports.

| *Year* | *Crude Materials* | *Crude Foodstuffs* | *Manufactured Food-stuffs* | *Semi Manfs.* | *Finished Manfs.* | *Miscellaneous* |
|---|---|---|---|---|---|---|
| 1821 | 60.46 | 4.79 | 19.51 | 9.42 | 5.66 | 0.16 |
| 1881 | 31.55 | 27.34 | 25.62 | 3.71 | 11.59 | 0.19 |
| 1924 | 29.49 | 8.73 | 12.75 | 13.57 | 35.32 | 0.14 |

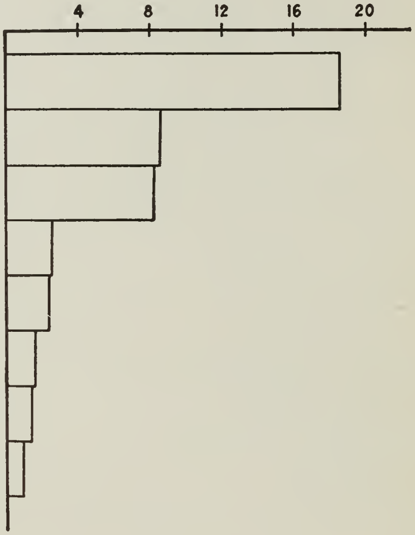
3. *p.* 583. Farm labor: proportion of each class to total for U. S.

| | *% U. S.* | *1917* |
|---|---|---|
| Hired by month: | | |
| *With board* | 36.1 | $28.87 |
| *Without board* | 15.5 | 40.43 |
| Day, excluding extra harvest: | | |
| *With board* | 15.3 | 1.56 |
| *Without board* | 15.7 | 2.02 |
| Day, harvest labor: | | |
| *With board* | 10.5 | 2.08 |
| *Without board* | 6.9 | 2.54 |

**30. Horizontal bar chart.**—Horizontal bar charts are used to represent a series of quantities where each item is a total in itself. Beginning at the left, one should enumerate the items in the first column. In a second column, place the numerical data. In a third column, place the horizontal bars. These bars should be the last column to the right on the page. The bars should begin at a uniform distance from a vertical base line. Bars should be of a uniform width, usually narrower than a hundred per cent bar. A scale should be placed above the bars. Thin lines should extend the scale across the field of the bars.

For historical data, the items should be placed in a time order. For geographical data, the items are placed in some order of contiguity. For the states of the United States, the order used by the Census Bureau is universally used. For many items, the arrangement is with respect to size. For some items, the most important is placed first. In the following table on membership in religious bodies the items are placed in order of size, the largest first. Consequently, we find, in fig. 3 the items are arranged according to size, the largest first.

| | |
|---|---|
| Roman Catholic................ | 18,260,793 |
| Methodist...................... | 8,433,268 |
| Baptist Bodies.................. | 8,189,448 |
| Presbyterians................... | 2,509,413 |
| Lutherans...................... | 2,465,841 |
| Disciples of Christ.............. | 1,383,247 |
| Protestant Episcopal............ | 1,128,859 |
| Congregational.................. | 857,846 |

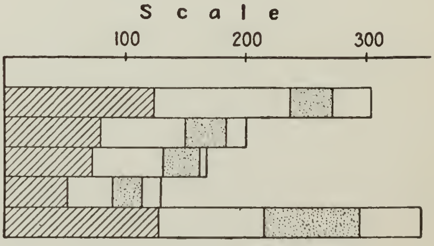


*Statistical Abstract of the U. S., 1924, p. 59.*

**Fig. 3.—Membership of Religious Bodies, 1923.**

**31. Compound bar chart.**—A compound bar chart is one in which each bar is a hundred per cent bar. The illustration in fig. 4 will make this clear.

| | *Manufacturing* | *Trading* | *Agents, Brokers, etc.* | *Banking* | *Grand Total* |
|---|---|---|---|---|---|
| 1913 | 123 | 115 | 34 | 32 | 304 |
| 1917 | 80 | 70 | 33 | 18 | 201 |
| 1918 | 73 | 58 | 32 | 5 | 168 |
| 1919 | 52 | 38 | 24 | 17 | 130 |
| 1920 | 128 | 89 | 79 | 51 | 346 |

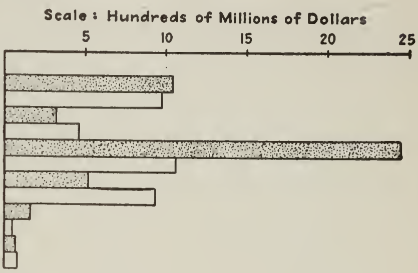


*Statistical Abstract of the U. S., 1924, p. 297.*

**Fig. 4.—Commercial Failures and Bank Suspensions. Liabilities in Millions of Dollars.**

**32. Multiple bar chart.**—If one desires to present two or more sets of data with respect to the same items, one constructs a multiple bar chart, as shown in fig. 5.

| | *Exports* (*000 omitted*) | *Imports* (*000 omitted*) |
|---|---|---|
| North America.. | $1,090,041 | $ 995,156 |
| South America.. | 314,252 | 466,074 |
| Europe......... | 2,445,300 | 1,096,087 |
| Asia............ | 514,592 | 930,708 |
| Oceania........ | 156,505 | 48,945 |
| Africa.......... | 70,294 | 72,992 |



*Statistical Abstract of the U. S., 1924, pp. 430–431.*

**Fig. 5.—Exports and Imports by Continents, 1924.**

**Exercises.**

Construct appropriate bar charts for the following data (*source: Statistical Abstract of the U. S., 1924*):

1. *p.* 33. Foreign white stock, Continental U. S., by mother tongue.

| | *1920* | *1910* |
|---|---|---|
| English and Celtic.................. | 9,729,365 | 9,930,861 |
| Germanic.......................... | 8,622,500 | 9,000,139 |
| Scandinavian...................... | 2,972,796 | 2,781,402 |
| Latin and Greek.................... | 6,036,001 | 4,185,932 |
| Slavic and Lettic.................. | 5,270,581 | 3,194,647 |
| Unclassified....................... | 2,956,321 | 2,261,563 |
| Unknown or mixed.................. | 811,394 | 888,838 |

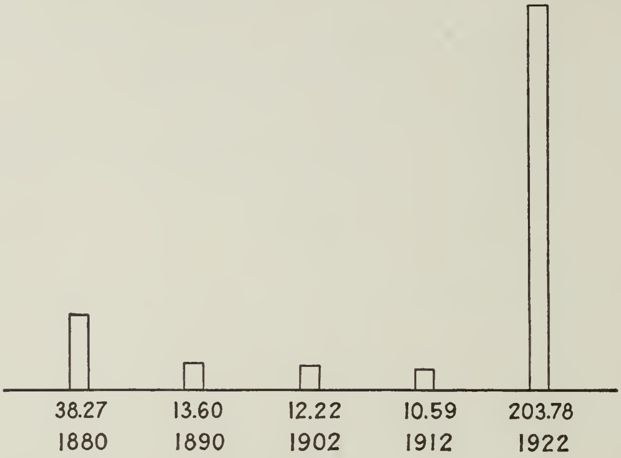
2. *p.* 441. Exports and imports, 1924 (000 omitted).

| | *Exports* | *Imports* |
|---|---|---|
| Atlantic Coast...................... | 2,245,602 | 2,357,722 |
| Gulf Coast......................... | 1,164,452 | 281,881 |
| Mexican Border.................... | 73,253 | 20,343 |
| Pacific Coast...................... | 447,311 | 477,302 |
| Northern Border................... | 638,946 | 441,717 |

3. *p.* 380. Persons killed and injured in railway accidents.

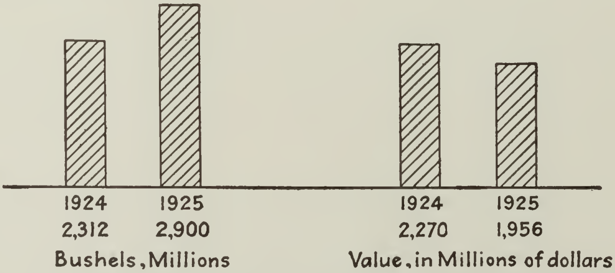
| | *Killed* | *Injured* |
|---|---|---|
| 1916................................ | 10,001 | 196,722 |
| 1918................................ | 9,286 | 174,575 |
| 1920................................ | 6,958 | 168,309 |
| 1922................................ | 6,325 | 134,871 |
| 1923................................ | 7,385 | 171,712 |
| 1924................................ | 6,617 | 143,739 |

**33. Vertical bar charts.**—In a vertical bar chart the bars are upright instead of horizontal. This is a natural viewpoint. We know instantly that it is the tops we must watch. The base line must not be omitted. Data and labels must be placed at the bottom of the columns. It becomes necessary in some cases to place the data and labels so that they are read from below upwards. These principles are illustrated by figs. 6 and 7.



*Statistical Abstract of U. S., 1924, p. 198.*

**Fig. 6.—Per Capita National Debt of the United States for Various Years, in Dollars.**



*Yearbook of Department of Agriculture, 1925, p. 788.*

**Fig. 7.—Corn, Production and Value, in the United States.**

The graph shows at a glance that the per capita debt decreased from 1880 to 1912. The World War was responsible for the decided increase in 1922.

Fig. 7 shows clearly that while the production of corn increased, the total value decreased. This is a better arrangement than to plot together the production and value for 1924 and again plot side by side the production and value for 1925.

## Exercises.

Construct vertical bar charts for the following data (*source: Statistical Abstract of the U. S., 1924*):

1. *p. 139.* Federal Civil Service employees.

| *Year* | *Number* | *Year* | *Number* |
|---|---|---|---|
| 1871 | 53,900 | 1911 | 370,000 |
| 1881 | 107,000 | 1921 | 597,482 |
| 1891 | 166,000 | 1922 | 560,863 |
| 1901 | 256,000 | 1923 | 548,531 |
| | | 1924 | 554,986 |

2. *p. 302.* Comparative international wholesale prices. (Index numbers for 1913 as a base. December, 1924.)

| | *U. S.* | *England* | *France* | *Canada* | *Japan* |
|---|---|---|---|---|---|
| On actual currency basis.... | 165 | 177 | 451 | 149 | 209 |
| Converted to gold basis..... | 165 | 171 | 126 | 149 | 161 |

3. *p. 308.* Retail food price index number, 1913 as a base.

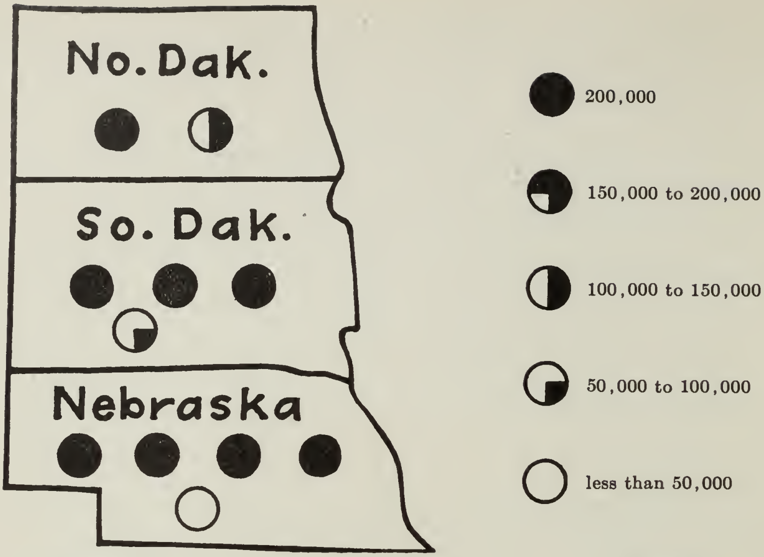
| *1890* | *1900* | *1910* | *1917* | *1918* | *1919* | *1920* | *1924* |
|---|---|---|---|---|---|---|---|
| 70 | 69 | 93 | 146 | 168 | 186 | 203 | 146 |

4. *p. 317.* Index numbers of union rates of wages per hour, 1913 base.

| *1907* | *1910* | *1917* | *1918* | *1919* | *1920* | *1924* |
|---|---|---|---|---|---|---|
| 90 | 94 | 114 | 133 | 155 | 199 | 228 |

5. *p. 567.* Average acreage per farm.

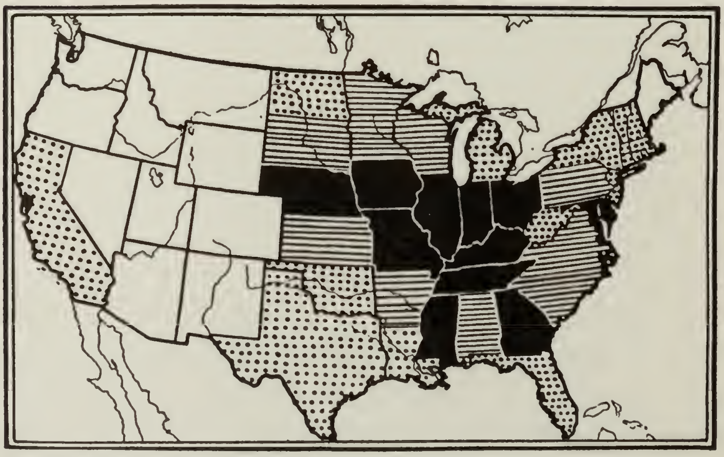
| *Year* | *Wyoming* | *California* | *Nebraska* | *Michigan* | *New York* | *Georgia* |
|---|---|---|---|---|---|---|
| 1910 | 777.6 | 316.7 | 297.8 | 91.5 | 102.2 | 92.6 |
| 1920 | 749.9 | 249.6 | 339.4 | 96.9 | 106.8 | 81.9 |



| North Dakota | South Dakota | Nebraska |
|---|---|---|
| 313,000 | 680,000 | 840,000 |

*Year Book of Department of Agriculture, 1925, p. 1148.*

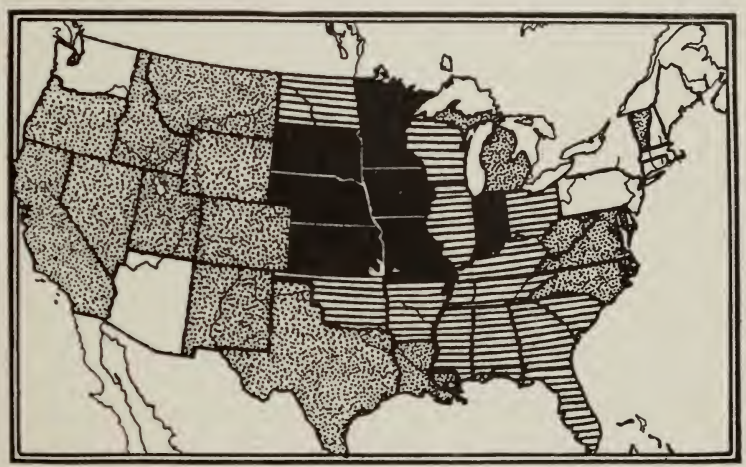
**Fig. 8.—Number of Sheep, Including Lambs, on Farms, Jan. 1, 1925.**



*1920 Census.*

Fig. 9.—Distribution of Swine Population. States shaded with black have 30 or more swine per square mile; those shaded with lines, 20–29; those dotted, 5–19; those unshaded, less than 5.

**34. Maps.**—The United States Census Bureau makes extensive use of maps in connection with the geographical distribution of items. Geographical maps are used to give the density in the distribution of items. Areas of varying density are distinguished by variations in shading, coloring, and cross-hatching. The density is sometimes indicated by the number of dots within the different areas. Fig. 8 illustrates these points. This kind of a map is sometimes referred to as a *dot diagram.*



*1920 Census.*

Fig. 10.—Per Capita Swine. States shaded with black have over eight-tenths head of swine per capita; those shaded with lines, one-half to eight-tenths; those dotted, two-tenths to one-half; those unshaded, less than two-tenths.

Fig. 9 gives the distribution of swine according to the 1920 census.

Fig. 10 gives the number of swine per capita of population.

**Exercises.**

Make diagrams for the following data (*source: Yearbook of the Department of Agriculture, 1925*):

1. All sheep on farms Jan. 1, 1925.

| | | | |
|---|---|---|---|
| *Wyoming* | 2,808,000 | *Utah* | 2,248,000 |
| *Montana* | 2,579,000 | *Colorado* | 2,616,000 |
| *Idaho* | 2,201,000 | *Nevada* | 1,108,000 |

2. *p. 1113.* Number of swine on farms Jan. 1, 1925.

| | | | |
|---|---|---|---|
| *Indiana* | 3,143,000 | *Ohio* | 2,421,000 |
| *Illinois* | 4,725,000 | *Nebraska* | 4,818,000 |
| *Iowa* | 9,633,000 | *Wisconsin* | 1,580,000 |
| *Kansas* | 2,467,000 | *Minnesota* | 3,600,000 |

3. *p. 1034.* Number of all cattle and calves on farms Jan. 1, 1925 (000 omitted).

| *Minn.* | *Iowa* | *Mo.* | *N. Dak.* | *S. Dak.* | *Neb.* | *Kans.* |
|---|---|---|---|---|---|---|
| 2,890 | 4,533 | 2,650 | 1,370 | 2,147 | 3,386 | 3,200 |

4. *p. 1036.* Cows and heifers 2 years old and over kept for milk Jan. 1, 1925 (000 omitted).

| *Minn.* | *Iowa* | *Mo.* | *N. Dak.* | *S. Dak.* | *Neb.* | *Kans.* |
|---|---|---|---|---|---|---|
| 1,563 | 1,341 | 835 | 520 | 544 | 625 | 766 |

**35. Pictorial diagrams.**[1]—As the name suggests, a pictorial diagram uses pictures of the thing under discussion. This is done in order that one may obtain not only an impression of the relative magnitudes involved, but also a knowledge of the nature of the item under discussion. Pictorial diagrams have no value for research. Their use is in attracting attention and popularizing a project.

This is a popular method of presentation, and is frequently used by newspapers and magazines. Pictorial diagrams may be used to popularize a campaign for a new high school building, a community chest budget, a national movement in support of aviation, or any undertaking of this nature.

In a pictorial diagram the bars are replaced by rows of pictures featuring the item under consideration. Thus we may have rows of men, the rows being of different length, to represent the changes in population. We may have rows

[1] Consult Karsten, "Charts and Graphs," Prentice-Hall, Inc., New York; and Brinton, "Graphic Methods," The Engineering Magazine Co., New York.

of boats to represent the marine tonnage in different years; rows of cars to represent the miles of railroad track in different years; rows of bales of cotton to represent the amounts of cotton in bales produced in different years; rows of cows to represent the varying number of milk cows per state.

In any given diagram all of the items should be of the same size, for it is the length of the rows that is being compared. To represent the changes in the number of milk cows, do not use two cows of different sizes to represent the number of cows under consideration. The observer might not thus obtain a proper mental picture. Shall the observer compare the lengths of the cows, the surface area, or the volume?

There are many pictorial diagrams in which all items are not of the same size. Such diagrams are generally to be condemned. For example, six men of different size are sometimes used to represent the relative sizes of the standing armies of the six leading nations. In what way did the illustrator intend that a comparison should be made? If comparison was meant to be on the basis of their heights, then those who compare surface area or volume obtain a false view. If comparison was meant to be on the basis of their surface area, then those who compare heights or volume obtain a false view. If comparison was meant to be on the basis of volume, then those who compare heights or surface area obtain a false view.

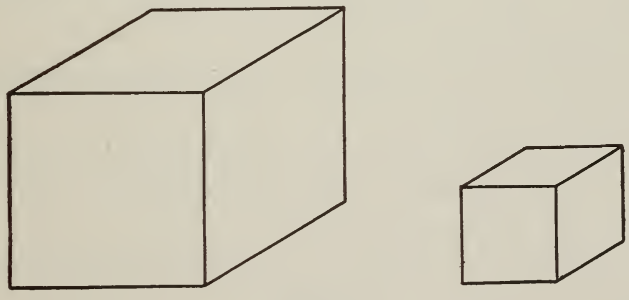


Fig. 11.

Here are two figures, the edge of the first being twice as large as that of the second. The area of the first is four times that of the second. The volume of the first is eight times that of the second. But the first does not give to the eye the impression of being eight times as large as the second. The eye does not estimate readily and accurately the comparative sizes of two similar three-dimensional bodies. There is added difficulty when the three-dimensional body is pictured on a surface. The eye does not estimate readily the comparative sizes of two similar two-space figures. Linear magnitudes are the most easily compared. Thus it is clear that all pictorial diagrams should be put in such form that linear magnitudes only are to be compared. If areas are present, they should be present in such form that one dimension (say, height) is the same throughout and hence the areas are to each other as their length. If volumes are present, they should be present in such form that they are to each other as some one linear dimension. Thus, in comparing standing armies, if lines are formed of a number of soldiers all of uniform size, then the volumes of the different lines are to each other as the lengths of the lines. Moreover, the surfaces are to each other as the lengths of the lines.

**36. Résumé**—*Choice of diagram.*—Given a particular set of data, the question arises as to what kind of a graph will best represent the data. This question has been answered in part in discussing the types of graphs. It is difficult to make sharp distinctions. In general, we state that:

1. A pie chart should be used for percentage distributions of a whole. This is especially true when the information refers to financial data.

2. Hundred per cent bars should be used for percentage distributions of a whole.

3. Horizontal bars should be used for simple comparisons of sizes of items especially where the division is made on the basis of attributes or categories.

4. Compound bar charts are simply hundred per cent horizontal bar charts.

5. Multiple bar charts should be used when one desires to present two or more sets of data with respect to the same items.

6. Vertical bar charts should be used in presenting frequency distributions and in graphing time series.

7. Maps are used in presenting density of a geographical distribution.

8. Pictorial diagrams are used to attract attention in a popular presentation.

## Chapter VI

# GRAPHIC REPRESENTATION—LINE GRAPHS

**37. Rectangular coördinate system.**—Draw two lines at right angles. Every point in the plane of the two lines can be located by giving its distance from each of the two lines. These two lines are called *axes of coördinates.* Their point of intersection is called the *origin of coördinates.* The line $OX$ in the figure is called the $x$-axis. The line $OY$ is called the $y$-axis. These two lines divide the region of the plane into four compartments called *quadrants.* These quadrants are numbered I, II, III, IV, counter-clockwise as shown. The distance of a point from the $y$-axis is termed an *abscissa.* The distance of a point from the $x$-axis is termed an *ordinate.* Abscissas measured to the right are positive; to the left, negative. Ordinates are positive when measured

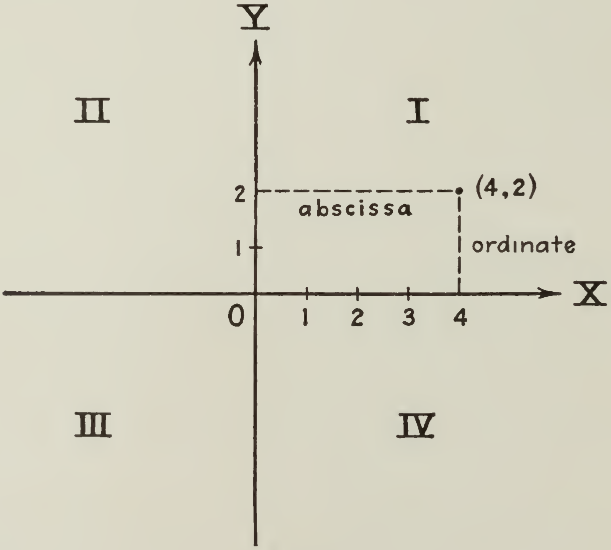


Fig. 12.

upward, negative when measured downward. The abscissa and ordinate together are called the *coördinates of a point.* In writing the coördinates, the convention is always to write the abscissa first.

**38. Line graphs.**—In constructing a graph of a *time* series it is customary to use times as abscissas and the variable size of the items as ordinates.

Table 20.—Population of the United States by Decades.

| *Year* | *Population* | *Year* | *Population* |
|---|---|---|---|
| 1850 | 23,191,876 | 1890 | 62,947,714 |
| 1860 | 31,443,321 | 1900 | 75,994,575 |
| 1870 | 38,558,371 | 1910 | 91,972,266 |
| 1880 | 50,155,783 | 1920 | 105,710,620 |

*Statistical Abstract of U. S., 1924, pp. 4–5.*

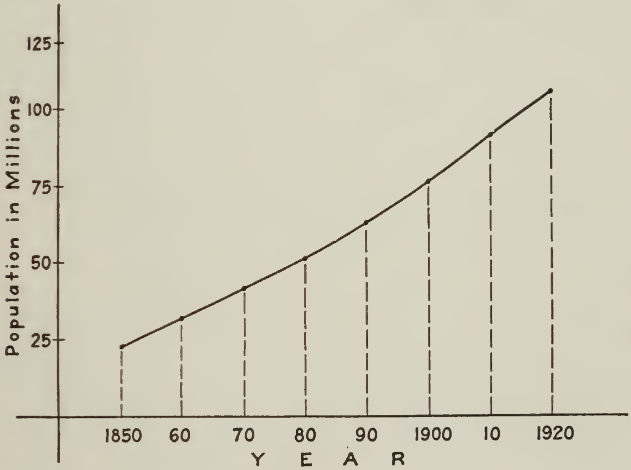


Fig. 13.—Population of the United States by Decades [See Table 20].

The year and the population for that year are the coördinates of a point. Plot all of the points as shown in fig. 13. Then join the points by as smooth a curve as possible. This curve represents graphically to the eye the manner of increase of population.

In constructing such a chart, a number of questions arise, the answers to which are embodied in the following general working rules.[2]

1. The general arrangement of a diagram should proceed from left to right.

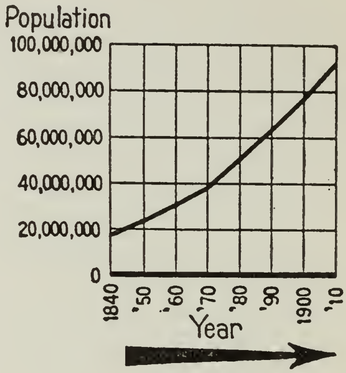


Fig. I.

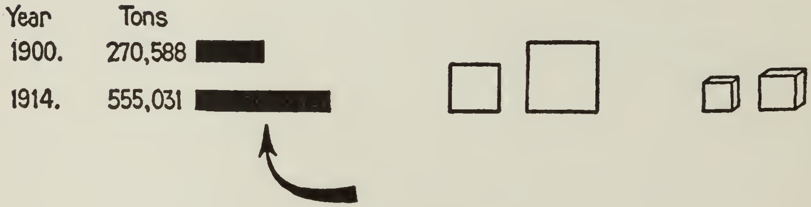


Fig. II.

2. Where possible represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

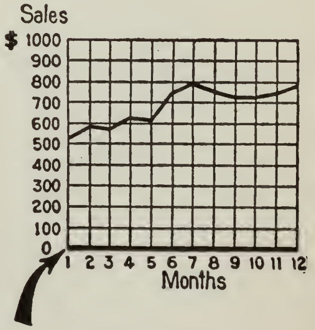


Fig. III.

[2] The report of the Joint Committee on Standards for Graphic Presentation, of which Mr. Willard C. Brinton was Chairman. Reproduced here by permission of The American Society of Mechanical Engineers, 29 West Thirty-ninth Street, New York.

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

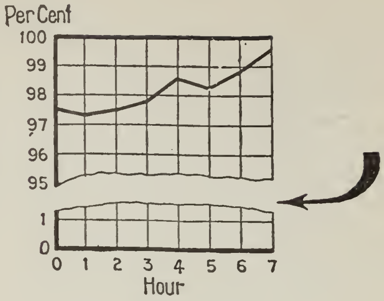


FIG. IV.

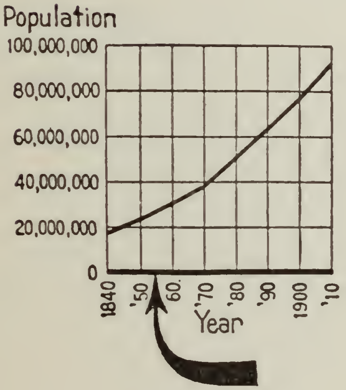


FIG. V–A.

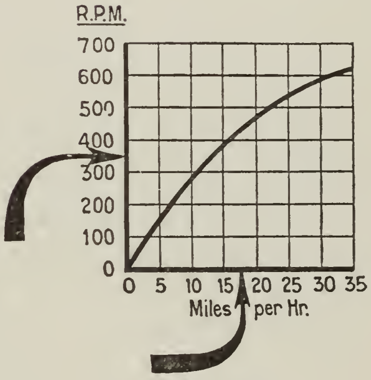


FIG. V–B.

5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

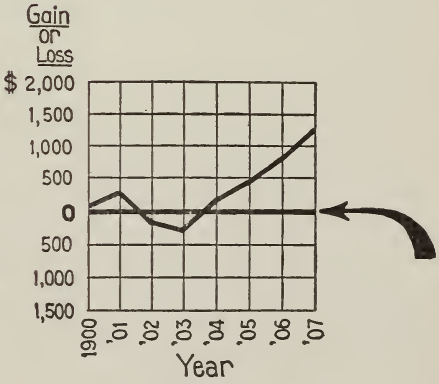


FIG. V–C.

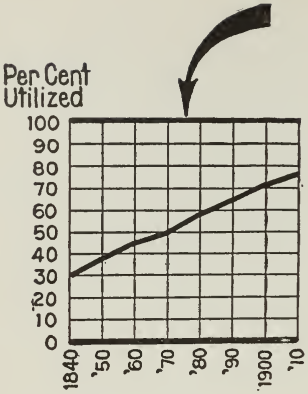


FIG. VI-A.

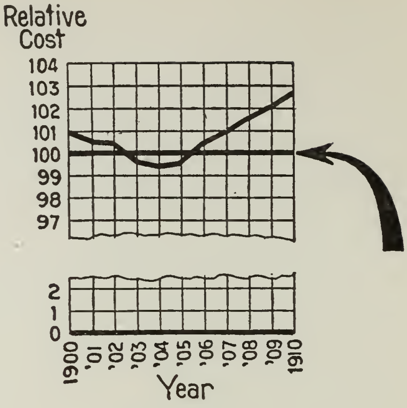


FIG. VI-B.

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

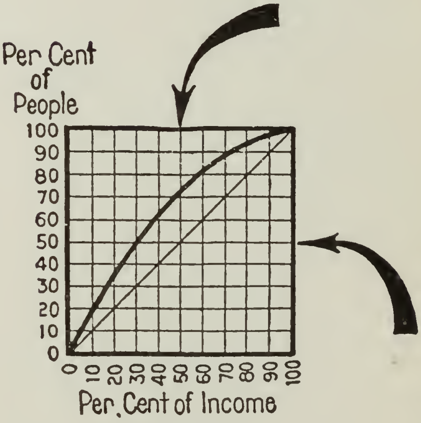


FIG. VI-C.

7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

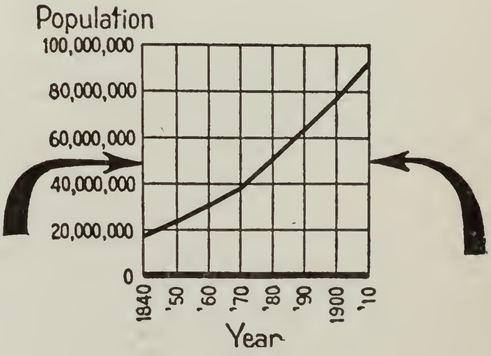


FIG. VII.

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

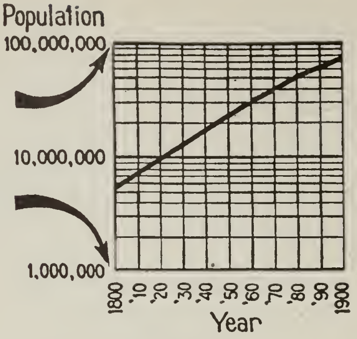


Fig. VIII.

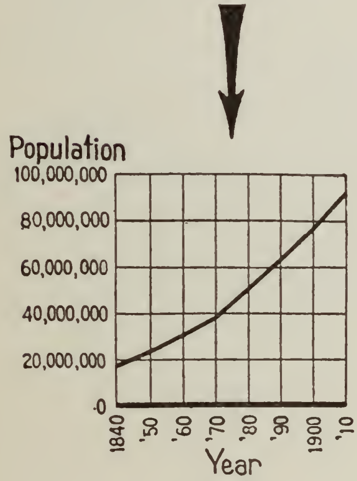


Fig. IX–A.

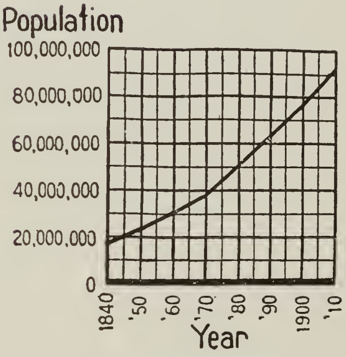


Fig. IX–B.

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10. The curve lines of a diagram should be sharply distinguished from the ruling.

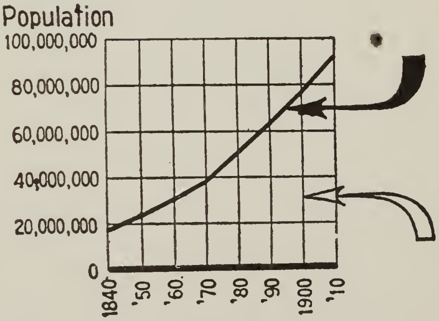


Fig. X.

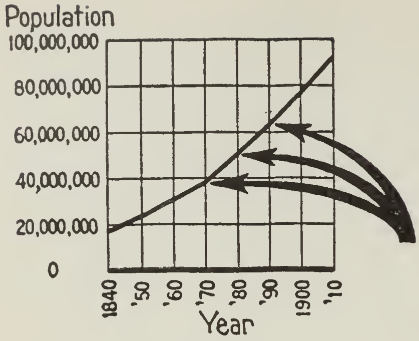


FIG. XI–A.

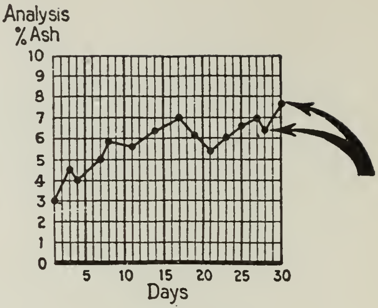


FIG. XI–B.

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

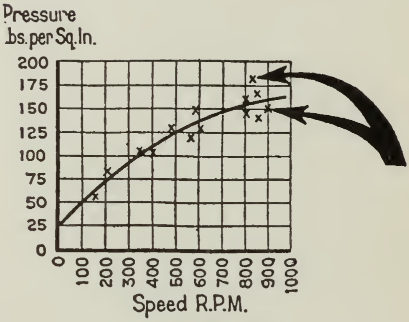


FIG. XI–C.

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

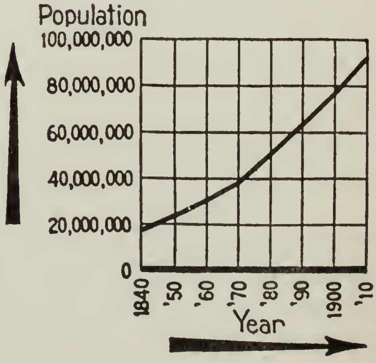


FIG. XII.

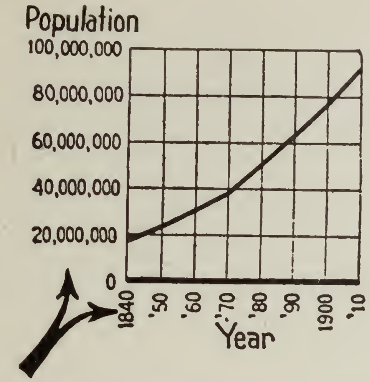


Fig. XIII-A.

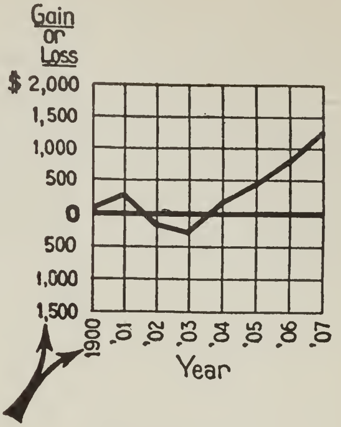


Fig. XIII-B.

13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

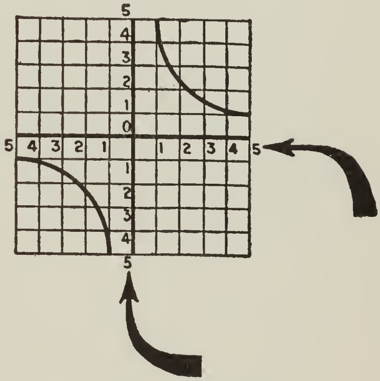


Fig. XIII-C.

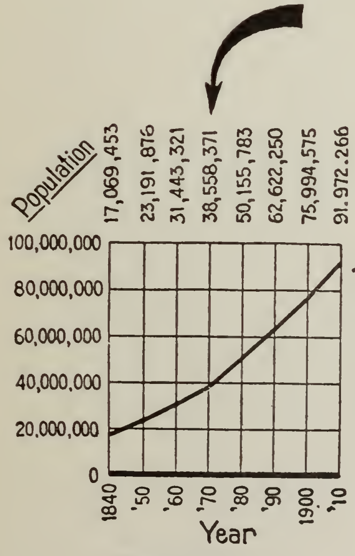


Fig. XIV-A.

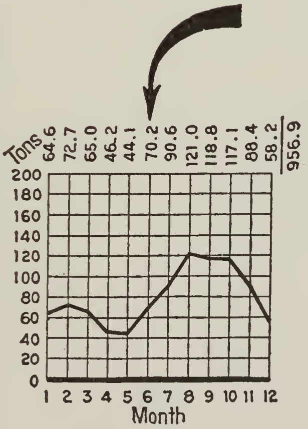


Fig. XIV-B.

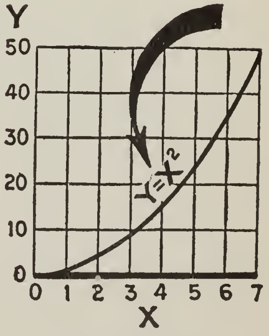
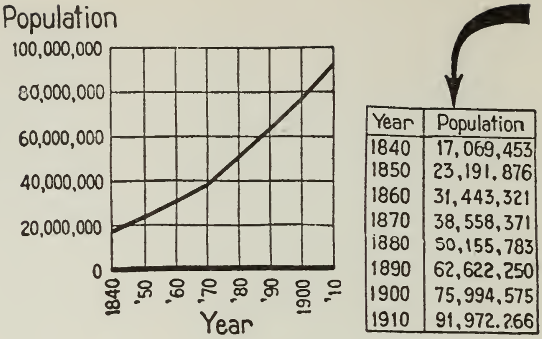


Fig. XIV-C.

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

15. If numerical data are not included in the diagram, it is desirable to give the data in tabular form accompanying the diagram.



| Year | Population |
|---|---|
| 1840 | 17,069,453 |
| 1850 | 23,191,876 |
| 1860 | 31,443,321 |
| 1870 | 38,558,371 |
| 1880 | 50,155,783 |
| 1890 | 62,622,250 |
| 1900 | 75,994,575 |
| 1910 | 91,972,266 |

Fig. XV.

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

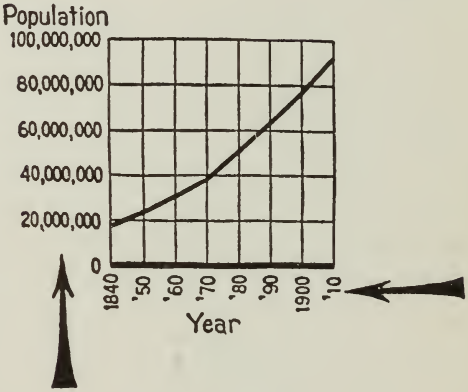


Fig. XVI.

17. The title of a diagram should be made as clear and complete as possible. Subtitles or descriptions should be added if necessary to insure clearness.

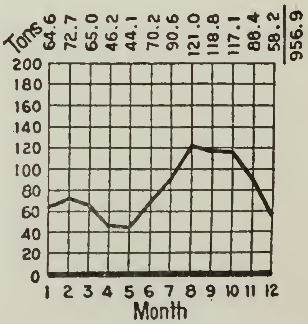


Fig. XVII.—Aluminum Castings Output of Plant No. 2, by Months, 1914. Output is given in short tons. Sales of Scrap Aluminum are not included.

**39. Unequal class intervals—Smoothing.**—A somewhat different procedure must be used in plotting the following data, in which the time intervals are not uniform.

**Table 21.—Deaths per 1,000 Births in the United States, during the First Year of Life, in 1923.**

| *Age* | *Deaths per 1,000 Births, 1923* |
|---|---|
| Under 1 mo. | 39.5 |
| 1 mo. | 6.4 |
| 2 mo. | 4.9 |
| 3 to 5 mo. | 11.2 |
| 6 to 8 mo. | 8.4 |
| 9 to 11 mo. | 6.8 |

*Statistical Abstract of U. S., 1924, p. 70.*

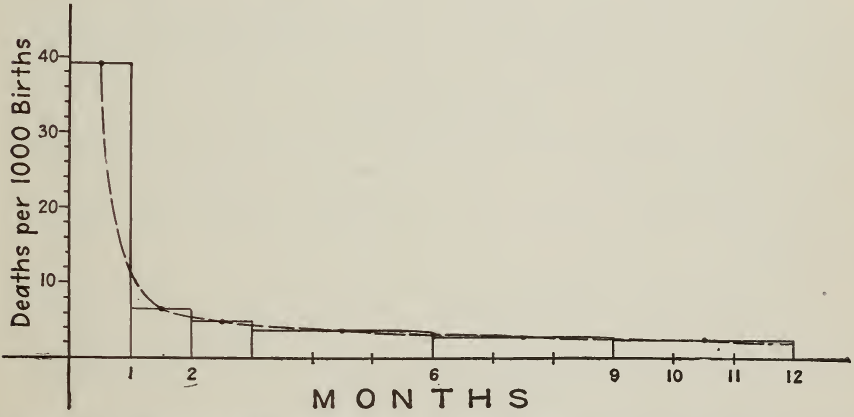


**Fig. 14.—Deaths per 1000 Births in the United States during the First Year of Life, in 1923.**

Construct a vertical bar chart. Use as base of the successive bars the successive age intervals. Construct on this base a rectangle whose area is equal to the designated number of deaths for that age interval. Thus, for the age interval 3 to 5 months the number of deaths recorded is 11.2. Construct a rectangular bar with a base of three units extending from 3 to 6. This base will represent the age interval. The altitude of this bar should be $\frac{11.2}{3} = 3.73$

units The area of this bar is then base times altitude, or $3 \times \frac{11.2}{3} = 11.2$, which number is the number of deaths recorded for the given age interval. The resulting diagram resembles a staircase. The various rectangles are called *frequency rectangles.* A curve called a *frequency polygon* is constructed by joining the mid-points of the tops of the successive rectangles. A *smoothed frequency curve* is obtained by joining the mid-points of the tops of the rectangles by a smooth curve drawn freehand. In drawing this smooth curve, the following rules should be observed as nearly as possible:

1. Keep the total area under the curve equal to the total area within the frequency rectangles.

2. Maintain individual areas. That is, the vertical sides of any rectangle should cut off under the curve an area equal to the area of the rectangle.

3. The highest point of the smoothed curve should extend above the highest point of the frequency polygon.

4. The curve must be free from sudden changes in direction.

**40. *Z*-charts.**—Plot on the same chart three graphs. One graph gives the production by months. A second graph gives the cumulative production. A third graph gives the moving yearly production. These three graphs form a letter *Z*. It is customary to use two different vertical scales, one for the monthly items and a smaller scale for the cumulative and moving totals. Such charts can be used to advantage in graphing production, sales, bank-clearings, total number of deaths, value of building construction, exports, and imports. Let us illustrate with the following data on value of construction in millions of dollars. (See page 59.)

The cumulative total for any month is obtained by adding the production for the current month to the cumulative total for the preceding month. In other words, the cumulative total for any month is the production for the

Table 22.—Value of Construction in Millions of Dollars by Months.

| *Months* | Value of Construction in Millions | | | Cumulative Total For | Moving Total For |
|---|---|---|---|---|---|
| | *1922* | *1923* | *1924* | *1923* | *1923* |
| January........... | 166 | 217 | 261 | 217 | 3,402 |
| February.......... | 177 | 230 | 259 | 447 | 3,455 |
| March............ | 294 | 334 | 386 | 781 | 3,495 |
| April............. | 353 | 357 | 426 | 1,138 | 3,499 |
| May.............. | 363 | 374 | 359 | 1,512 | 3,510 |
| June.............. | 343 | 324 | 331 | 1,836 | 3,491 |
| July.............. | 350 | 274 | 290 | 2,110 | 3,415 |
| August............ | 322 | 253 | 300 | 2,363 | 3,346 |
| September......... | 271 | 254 | 298 | 2,617 | 3,329 |
| October.......... | 253 | 320 | 345 | 2,937 | 3,396 |
| November......... | 244 | 289 | 341 | 3,226 | 3,441 |
| December......... | 215 | 268 | 283 | 3,494 | 3,494 |

*Statistical Abstract of U. S., 1924, p. 788.*

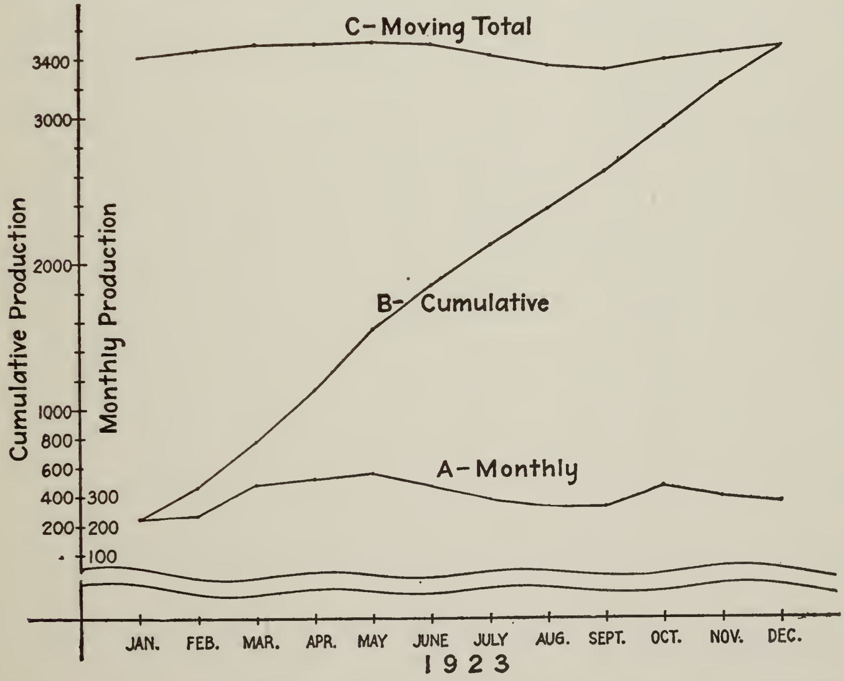


Fig. 15.—Chart Showing Value of Construction in the United States in Millions of Dollars, for 1923: A, by Months; B, Cumulative; C, Moving Total.

year beginning with January and ending with the current month. Thus the cumulative total for March, 1923, is:

$$217 + 230 + 334 = 781$$

The moving total is the total production for a twelve-months' period ending with the current month. Thus the totals for December in the last two columns must agree.

### Exercises.

Plot the following data:

1. Revenue in millions of dollars, ton-miles in trillions, 1924.

| | *Jan.* | *Feb.* | *Mar.* | *Apr.* | *May* | *June* | *July* | *Aug.* | *Sept.* | *Oct.* | *Nov.* | *Dec.* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Freight revenue... | 333.6 | 352.4 | 371.6 | 342.8 | 344.7 | 323.3 | 339.9 | 358.4 | 399.0 | 438.8 | 380.9 | 362.4 |
| Passenger revenue. | 91.7 | 83.4 | 87.1 | 85.2 | 85.6 | 96.0 | 97.4 | 104.5 | 93.2 | 82.9 | 78.8 | 90.8 |
| Freight ton-miles.. | 34.5 | 36.0 | 36.4 | 31.9 | 33.9 | 31.9 | 33.1 | 36.4 | 39.0 | 43.1 | 38.0 | 35.0 |

*Statistical Abstract of the U. S., 1924, p. 368.*

2. Deaths by Age in the Registration Area for 1924.

| *Age* | *Typhoid and Paratyphoid Fever* | *Smallpox* | *Measles* | *Scarlet Fever* | *Whooping Cough* | *Erysipelas* | *Dysentery* | *Rickets* | *Chronic Rheumatism and Gout* |
|---|---|---|---|---|---|---|---|---|---|
| Under 1 yr...... | 48 | 63 | 2,024 | 131 | 4,844 | 772 | 797 | 346 | 1 |
| 1 yr...... | 62 | 15 | 2,343 | 301 | 1,922 | 48 | 541 | 120 | 1 |
| 2 yr...... | 66 | 7 | 883 | 346 | 623 | 27 | 163 | 25 | |
| 3 yr...... | 70 | 6 | 506 | 334 | 290 | 15 | 79 | 14 | |
| 4 yr...... | 75 | 13 | 319 | 282 | 158 | 2 | 27 | 3 | |
| 5–9........... | 470 | 26 | 900 | 803 | 258 | 17 | 51 | 12 | 6 |
| 10–14.......... | 717 | 8 | 345 | 278 | 42 | 21 | 19 | 4 | 5 |
| 15–19.......... | 1,155 | 47 | 336 | 167 | 16 | 31 | 27 | 4 | 14 |
| 20–24.......... | 1,020 | 92 | 154 | 151 | 5 | 45 | 40 | | 14 |
| 25–29.......... | 615 | 105 | 95 | 108 | 5 | 55 | 38 | 1 | 33 |
| 30–34.......... | 460 | 63 | 90 | 94 | | 52 | 40 | | 22 |
| 35–44.......... | 797 | 165 | 180 | 74 | 7 | 193 | 107 | | 80 |
| 45–54.......... | 587 | 109 | 116 | 34 | 1 | 253 | 101 | | 137 |
| 55–64.......... | 308 | 81 | 88 | 9 | 3 | 293 | 153 | | 232 |
| 65–74.......... | 170 | 47 | 80 | 6 | 3 | 314 | 278 | | 359 |
| 75 and over..... | 46 | 27 | 47 | 2 | 5 | 319 | 473 | | 371 |
| Unknown....... | 11 | | 11 | 2 | 6 | 1 | 12 | | 3 |
| TOTAL........ | 6,677 | 874 | 8,517 | 3,122 | 8,188 | 2,458 | 2,946 | 529 | 1,278 |

*Bureau of Census, Mortality Statistics, 1924, p. 223.*

Plot the following data, which are to be found in the *Statistical Abstract of the U. S. for 1924.*

3. *p. 61. Table no. 50.* Deaths. Rate per 1,000 population.

4. *p. 102. Table no. 97.* Vocational Education. Total teachers and pupils in vocational schools for the years tabulated.

5. *p. 134. Table no. 125.* Total pensioners on rolls.

6. *p. 139. Table no. 135.* Approximate number of employees in the executive civil service and the merit system.

7. *p. 65. Table no. 55.* Deaths. Rates per 100,000 population by important causes for the years tabulated.

8. *p. 787. Table no. 725.* Index numbers of building materials and construction costs, 1913–1924 inclusive.

9. Plot the curves for the average farm prices, Dec. 1, of the following agricultural products. *Data are found in the Yearbook of the Department of Agriculture for the years specified and at the page designated.*

| | | | |
|---|---|---|---|
| (a) Wheat | 1917, p. 615 | (b) Corn | 1901, p. 699 |
| | 1922, p. 583 | | 1922, p. 571 |
| | 1925, p. 779 | | 1925, p. 788 |
| (c) Oats | 1901, p. 718 | (d) Barley | 1901, p. 727 |
| | 1922, p. 620 | | 1922, p. 631 |
| | 1925, p. 806 | | 1925, p. 821 |
| (e) Potatoes | 1901, p. 741 | (f) Cotton | 1901, p. 754 |
| | 1922, p. 668 | | 1922, p. 711 |
| | 1925, p. 913 | | 1925, p. 952 |

10. Plot a *Z*-chart for the year 1924, using the data in table 22.

**41. Frequency distributions.**—For some distributions we are interested in the numerical magnitude of a stated characteristic and the frequency with which any magnitude occurs. The time element does not enter. We are interested in making a column diagram, frequency polygon, or smoothed line curve to represent such a distribution. Usually the measured magnitude is plotted as a horizontal abscissa with the corresponding frequencies plotted vertically as ordinates. These frequency curves usually take on the shape of one of the following:

(a) Symmetric or bell-shaped curve.
(b) *J*-shaped curve.
(c) *U*-shaped curve.
(d) Asymmetric or skewed curve.

**42. Bell-shaped curve.**—Curves of this type are met in frequency distributions of biological data. The curve has a high point in the middle and drops off uniformly on either side. A characteristic measured for a number of items will show a given magnitude more frequently than any other magnitude. Measures differing from this most frequent magnitude by the same amount, more or less, are found to occur with about the same frequency.

One obtains the ideal bell-shaped curve in plotting the theoretical frequencies of the number of heads occurring in 128 throws with 6 pennies.

**Table 23-A.—Frequencies of Heads in 128 Throws with 6 Coins.**

| *No. of heads* | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| *Frequencies* | 2 | 12 | 30 | 40 | 30 | 12 | 2 |

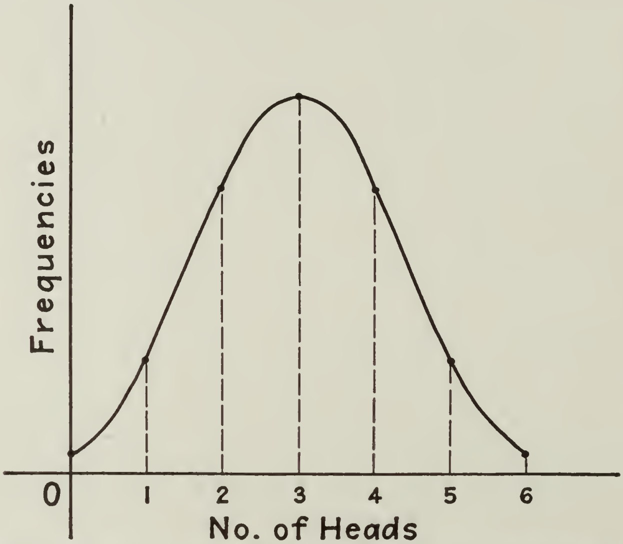


**Fig. 16.—Frequencies of Heads in 128 Throws with Six Coins.**

In an actual trial of 128 throws, the frequencies would probably differ slightly from the theoretical frequencies and the corresponding curve would not be absolutely symmetrical.

The following data[1] give the frequencies of weights in ounces of ears of corn.

**Table 23-B.—Frequencies of Weights in Ounces of 993 Ears of Corn.**

| *Weight, oz.* | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Frequencies* | 4 | 22 | 27 | 50 | 47 | 71 | 75 | 71 | 75 | 88 | 107 | 114 | 112 | 65 | 37 | 8 | 13 | 4 | 2 | 1 |

Total 993.

[1] E. Davenport, "Principles of Breeding," p. 461. Ginn & Co.

The frequency polygon here constructed shows a number of irregularities. These irregularities are due to the sample. Another sample would have other irregularities but would have the peak at about the same place and would taper off on either side in the same general way.

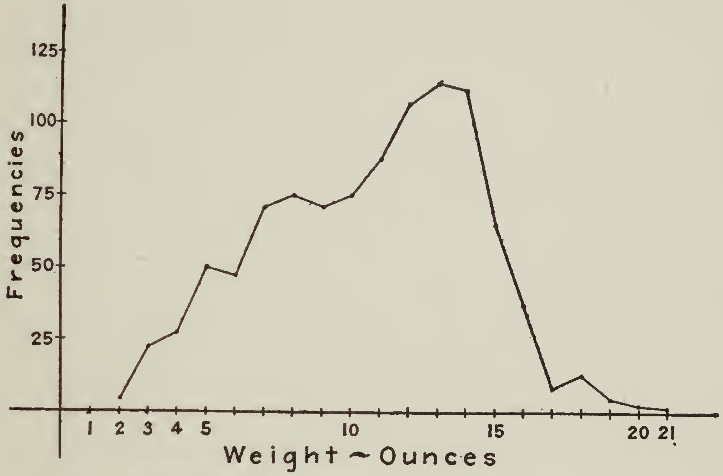


**Fig. 17.—Frequencies of Weights in Ounces of 993 Ears of Corn.**

Biological measurements, such as stature, weight, and intellectual capacity, are distributed according to the normal law, giving a fairly symmetrical bell-shaped curve.

But are economic[2] measurements, such as production of pig iron, interest rate on sixty-to-ninety day paper, and value of building permits distributed according to the same normal law? Table 23-C shows the frequency distribution of pig iron and interest rate on sixty-to-ninety day paper. For narrow class intervals . . . the production of pig iron reveals a bimodal distribution; for wider class intervals the distribution is approximately normal. The same type of distribution, though less marked, is revealed by the . . . interest rate on sixty-to-ninety day paper; in this case the distribution is . . . unsymmetrical. . . . Theoretically, the *irregular* fluctuations rather than the *cyclical* fluctuations are the ones which we should expect to be distributed according to the normal law of distribution of errors.

[2] Quoted from "Indices of General Business Conditions" by W. M. Persons, Harvard University Press, Cambridge, Mass., 1919, p. 137.

The *irregular* fluctuations are secured by eliminating secular trend, cyclical and seasonal fluctuations from the original data.[3]

**Table 23-C.—Frequency Table of (a) Interest Rates on 60–90 Day Paper July, 1903–July, 1914, and (b) Monthly Production of Pig Iron, Jan., 1903–Jan., 1914 (Class interval 0.3$\sigma$).**

| Measure Above or Below Mean | Frequency | | Measure Above or Below Mean | Frequency | |
|---|---|---|---|---|---|
| | *Interest Rate on 60–90 Day Paper* | *Production of Pig Iron* | | *Interest Rate* | *Production of Pig Iron* |
| −2.4 | 0 | 3 | +0.3 | 11 | 13 |
| --2.1 | 0 | 5 | 0.6 | 10 | 22 |
| −1.8 | 0 | 2 | 0.9 | 16 | 23 |
| −1.5 | 7 | 4 | 1.2 | 11 | 12 |
| −1.2 | 9 | 3 | 1.5 | 5 | 3 |
| −0.9 | 17 | 11 | 1.8 | 4 | 0 |
| −0.6 | 15 | 15 | 2.1 | 2 | 0 |
| −0.3 | 11 | 9 | 2.4 | 1 | 0 |
| 0 | 13 | 8 | 2.7 | 1 | 0 |

### Exercises.

1. Plot the data in table 23-C.

2. Make a new frequency distribution from the data in table 23-C, using as class interval, 0.9$\sigma$.

3. Plot the frequency distributions obtained in exercise 2.

4. Plot the frequency distribution in table VI, appendix.

5. Plot the frequency distribution in table XIX, appendix.

**43. *J*-shaped curve.**—This curve is formed like the letter *J*. The greatest frequency is at one end or the other of the distribution. Some *J*-shaped curves would lose this shape if the class intervals were made smaller. A curve of frequencies of leaves on clover would be a true *J*-shaped

[3] For a method of making this elimination, consult "Indices of General Business Conditions," by W. M. Persons, Harvard University Press, Cambridge, Mass., 1919. pp. 137–138.

curve, for the frequency is never less than three and this is by far the greatest frequency.

We give, in fig. 18, a *J*-shaped curve for the frequencies of different numbers of petals for three series of *Ranunculus bulbosus.*

**Table 24.—Frequencies of Different Numbers of Petals for Three Series of Ranunculus Bulbosus.**

| *Number of Petals* | Series I | II | III |
|---|---|---|---|
| 5 | 312 | 345 | 133 |
| 6 | 17 | 24 | 55 |
| 7 | 4 | 7 | 23 |
| 8 | 2 | ... | 7 |
| 9 | 2 | 2 | 2 |
| 10 | ... | ... | 2 |
| 11 | ... | 2 | |
| Total.......... | 337 | 380 | 222 |

*H. De Vries, Br. Dtsch. Bot. Ges., Bd. XII, 1894.*

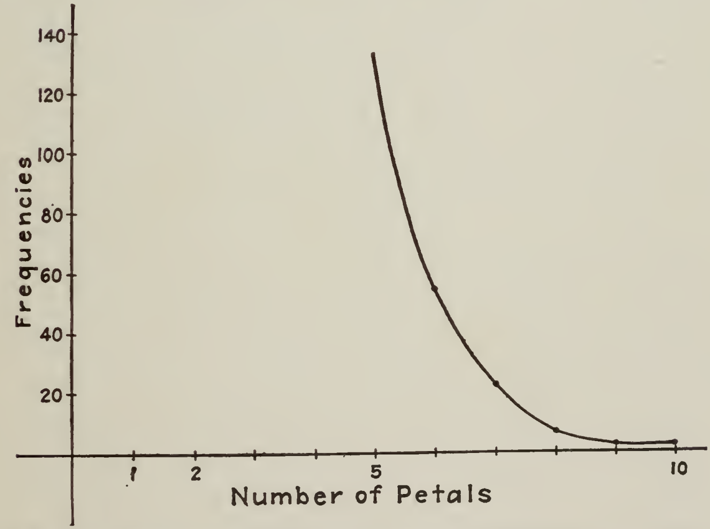


**Fig. 18.—Frequencies of Petals of Ranunculus Bulbosus. Series III (Table 24).**

A *J*-shaped curve is the typical curve of forgetting, when days are used as abscissa and the per cent retained as ordinates. That is, the amount retained decreases with increase of time. For such a curve consult A. J. Snow, "Psychology in Business Relations," p. 309, fig. 29.

The data in ex. 2., p. 60, on deaths by age due to whooping-cough and rickets give *J*-shaped curves. The number of men unmarried at given ages must give a true *J*-shaped curve. See table XX, appendix. From the same table, the death-rate gives also a *J*-shaped curve. Table XIV, appendix, would give a *J*-shaped curve if the interval were $2,000 instead of $1,000. This distribution is not essentially *J*-shaped but extremely asymmetrical.

**44. *U*-shaped curve.**—The *U*-shaped curve is not common. Yule in his "Introduction to the Theory of Statistics," p. 104, says that the *U*-shaped form of distribution appears sometimes to be exhibited by the percentage of offspring possessing a certain attribute when at least one of the parents also possesses the attribute.

| *Percentage Deaf-mutes* | *Number of Families* | *Percentage Deaf-mutes* | *Number of Families* |
|---|---|---|---|
| 0–20 | 220 | 60– 80 | 5.5 |
| 20–40 | 20.5 | 80–100 | 15.0 |
| 40–60 | 12 | | |

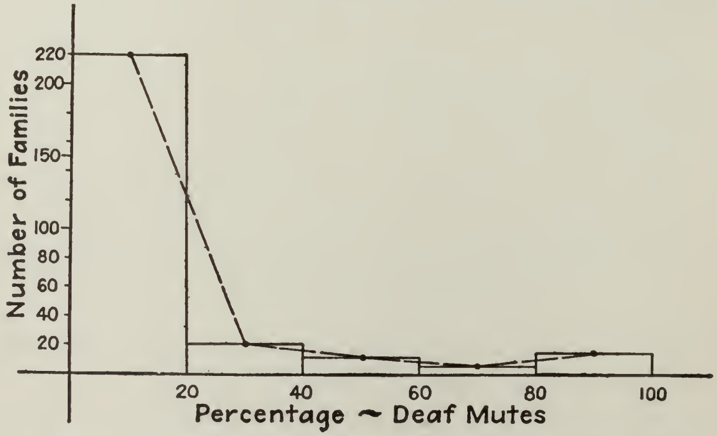


**Fig. 19.**—*U*-shaped Curve Showing Number of Families Having the Given Percentage of Deaf-Mutes.

The stock example of this type of curve is the degree of cloudiness of the sky observed at Breslau during the years 1876–1885, on a scale of 10.

**Table 25.—Number of Days with Stated Degree of Cloudiness at Breslau during the Years 1876–1885, on a Scale of 10.**

| *Degrees* | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *No. of days* | 751 | 179 | 107 | 69 | 46 | 9 | 21 | 71 | 194 | 117 | 2,086 |

*Yule, p. 103.*

It is obvious that the graph of these data would form a *U*-shaped curve.

A *U*-shaped curve is obtained by using minutes as ordinates and, as abscissa, diastolic blood pressure in mm. following an intravenous injection of sodium Iodoxybenzoate in Arthritis. For such a curve consult the *Journal of the American Medical Association,* vol. 89, no. 14, Oct. 1, 1927, p. 1125. There is uniformly a sharp decline in blood pressure below preinjection pressure, followed by a gradual rise to normal, requiring from one to five hours.

The data in ex. 2, p. 60, on deaths by age due to Erysipelas and Dysentery give *U*-shaped curves.

The general death-rate per 1,000 population gives a *U*-shaped curve for the death-rate in early youth and in old age is greater than in middle life.

**45. Skewed curves.**—A bell-shaped curve is said to be positively skewed when it has a long tail on the right. A curve is negatively skewed when it has a long tail to the left. There is a tendency to positive skewness when the range is limited on the left and unlimited on the right. This happens for income data and for index numbers.

This matter is illustrated by fig. 20, constructed from data in table 26. (See p. 68.)

For the 16 miners who received a wage of 30¢ or less, we have no data to show the minimum wage received. It would seem that we are forced to consider zero as the minimum and draw a rectangle whose base is three units (30¢)

Table 26.—Miners, Hand or Pick, Colcrado, at Face, 1924.

| *Wages per Hour* | *Frequency* | *Wages per Hour* | *Frequency* |
|---|---|---|---|
| under 30¢... | 16 | 1.30 and under 1.40..... | 44 |
| 30 and under 40.... | 21 | 1.40 " " 1.50..... | 39 |
| 40 " " 50.... | 42 | 1.50 " " 1.60..... | 20 |
| 50 " " 60.... | 53 | 1.60 " " 1.70..... | 20 |
| 60 " " 70.... | 104 | 1.70 " " 1.80..... | 12 |
| 70 " " 80.... | 143 | 1.80 " " 1.90..... | 13 |
| 80 " " 90.... | 143 | 1.90 " " 2.00..... | 9 |
| 90 " " 1.00.... | 155 | 2.00 " " 2.50..... | 5 |
| 1.00 " " 1.10.... | 141 | 2.50 " " 3.00..... | 4 |
| 1.10 " " 1.20.... | 94 | 3.00 " over.......... | 1 |
| 1.20 " " 1.30.... | 67 | | |
| | | Total................ | 1,146 |

*Bulletin No. 416 of the U. S. Bureau of Labor Statistics, p. 55.*

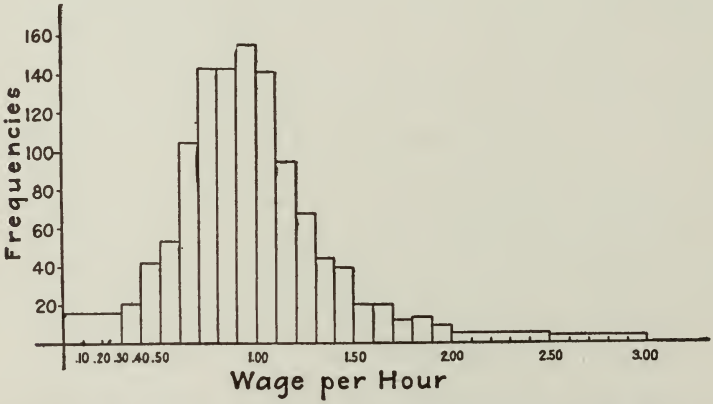


Fig. 20.—Skewed Diagram, Showing Frequency with Which Wages in Cents per Hour Were Paid to Miners (Hand or Pick at Face) in Colorado, 1924. (Data, Table 26.)

and altitude sufficient to make the area 16 units. There are five workers who receive a wage from $2.00 to $2.50. We construct a rectangle with a base of five units (50¢) and altitude sufficient to make the total area five. We have no means of knowing what altitude or base to assign for the one worker whose wage is $3.00 or over. A smoothed frequency curve would obviously have a long tail on the right.

## Exercises.

Plot the following data:

1. Distribution of grades of Colorado College freshmen, June, 1923.

| *Subject* | *A* | *B* | *C* | *D* | *E* | *F* | Total |
|---|---|---|---|---|---|---|---|
| Trigonometry.......... | 24 | 42 | 24 | 30 | 1 | 6 | 127 |
| English............... | 23 | 40 | 44 | 18 | 2 | 0 | 127 |
| Biology............... | 18 | 63 | 26 | 14 | 1 | 2 | 124 |
| History 1A............ | 6 | 20 | 37 | 19 | 7 | 1 | 90 |
| Biblical Lit........... | 18 | 25 | 6 | 0 | 0 | 2 | 51 |
| French................ | 6 | 9 | 15 | 8 | 0 | 2 | 40 |

2. Estimated distribution of income among the single women of the continental United States in 1910. (*Source: King, "Wealth and Income" p. 224.*)

| *Income $* | *No.* | *Income $* | *No.* | *Income $* | *No.* |
|---|---|---|---|---|---|
| 0–200 | 10 | 600– 700 | 150 | 1,100–1,200 | 12 |
| 200–300 | 70 | 700– 800 | 110 | 1,200–1,300 | 8 |
| 300–400 | 560 | 800– 900 | 37 | 1,300–1,400 | 5 |
| 400–500 | 530 | 900–1,000 | 22 | | |
| 500–600 | 280 | 1,000–1,100 | 16 | Total.......... | 1,810 |

3. Personal income tax returns by income classes for 1922. (*Statistical Abstract of the U. S., 1924, p. 162.*)

| *Income Class* | *No. of Returns* | *Income Class* | *No. of Returns* |
|---|---|---|---|
| Under $1,000 | 402,076 | 50,000 to 100,000 | 12,000 |
| 1,000 to 2,000 | 2,471,181 | 100,000 to 150,000 | 2,171 |
| 2,000 to 3,000 | 2,129,898 | 150,000 to 300,000 | 1,323 |
| 3,000 to 5,000 | 1,190,115 | 300,000 to 500,000 | 309 |
| 5,000 to 10,000 | 391,373 | 500,000 to 1,000,000 | 161 |
| 10,000 to 25,000 | 151,329 | 1,000,000 and over | 67 |
| 25,000 to 50,000 | 35,478 | Total.................. | 6,787,481 |

4. Tax rate on net income by income classes for 1922. (*Statistical Abstract of the U. S., 1924, p. 162.*)

| *Income Class* | *Tax Rate %* | *Income Class* | *Tax Rate %* |
|---|---|---|---|
| Under $1,000 | 0.10 | 50,000 to 100,000 | 17.89 |
| 1,000 to 2,000 | 0.75 | 100,000 to 150,000 | 27.42 |
| 2,000 to 3,000 | 0.40 | 150,000 to 300,000 | 37.03 |
| 3,000 to 5,000 | 1.06 | 300,000 to 500,000 | 37.27 |
| 5,000 to 10,000 | 2.66 | 500,000 to 1,000,000 | 35.81 |
| 10,000 to 25,000 | 5.48 | 1,000,000 and over | 35.02 |
| 25,000 to 50,000 | 10.40 | Average rate............ | 4.04 |

5. United States merchant marine: vessels by sizes, as of July 1, 1924· (*Statistical Abstract of the U. S., 1924, p. 394.*)

| *Gross Tons* | *Number* | *Gross Tons* | *Number* |
|---|---|---|---|
| 1,000 to 1,999............ | 165 | 6,000 to 6,999.......... | 350 |
| 2,000 to 2,999............ | 630 | 7,000 to 7,999.......... | 153 |
| 3,000 to 3,999............ | 306 | 8,000 to 8,999.......... | 58 |
| 4,000 to 4,999............ | 230 | 9,000 to 9,999.......... | 23 |
| 5,000 to 5,999............ | 503 | 10,000 and over.......... | 43 |

6. *Statistical Abstract of the U. S., 1924.*

(a) *p. 565.* All farm land, 1920, acreage by size of farms; also per cent distribution of acreage by size of farms.

(b) *p. 683.* Size of producing establishments, mines, and quarries, by value of products and by average number of wage earners. (1) All industries; (2) anthracite coal; (3) bituminous coal; (4) petroleum; (5) iron ore.

(c) *p. 724.* Manufactures. (1) Number of establishments classified according to average number of wage earners, (2) number of establishments classified by value of products.

7. Per cent of the total enrollment in public schools in the United States enrolled in each grade during the school year 1923–1924. (*Source: Survey of Education in Utah*):

| *Grades......* | K | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Per cent....* | 2.6 | 17.3 | 11.7 | 11.2 | 11.1 | 10.1 | 8.8 | 7.6 | 5.7 | 5.5 | 3.9 | 2.8 | 1.7 |

**46. Zones.**—The prices of stocks and bonds fluctuate. It is impracticable to plot the daily prices for a series of years. There is a variance, almost every day, between the high and low price. What is done in plotting the price of a stock or bond for a series of years is to plot two curves; one for the high price in each year, and one for the low price

**Table 27.—Yearly Price Range of Common Stock in the American Telephone and Telegraph Company from 1905 to 1926.**

| Year | *Low* | *High* | Year | *Low* | *High* |
|---|---|---|---|---|---|
| 1905 | 131 | 148½ | 1915 | 116 | 130¼ |
| 06 | 130 | 144⅝ | 16 | 123⅛ | 134⅛ |
| 07 | 88 | 133 | 17 | 95¾ | 128½ |
| 08 | 101 | 132⅝ | 18 | 90⅝ | 109¼ |
| 09 | 125 | 145⅛ | 19 | 95 | 108⅝ |
| 1910 | 126¾ | 143⅜ | 1920 | 92⅛ | 100¾ |
| 11 | 131½ | 153⅛ | 21 | 95¾ | 108½ |
| 12 | 137⅝ | 159⅛ | 22 | 114½ | 128¼ |
| 13 | 110 | 130 | 23 | 119⅛ | 128¾ |
| 14 | 114 | 124¼ | 24 | 121⅛ | 134¾ |
| | | | 25 | 130⅝ | 145 |
| | | | 26 | 139⅝ | 151 |

*Poor's Manual of Railroads, Commercial and Financial Chronicle.*

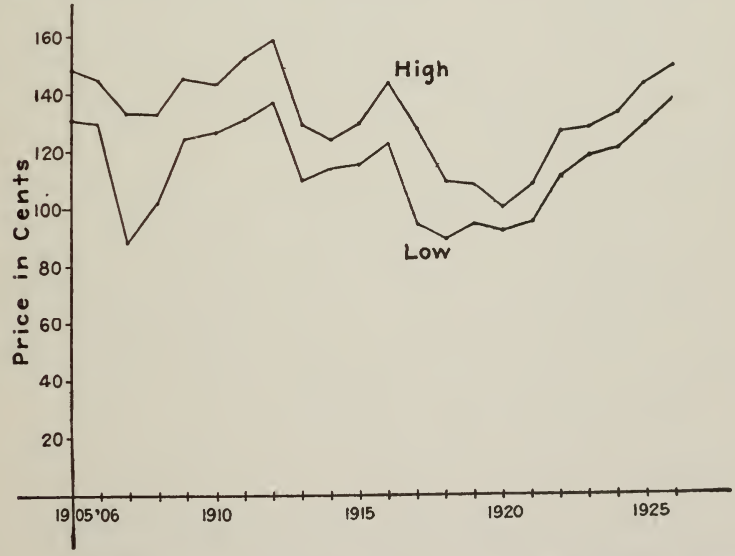


Fig. 21.—Yearly Price Range of Common Stock in the American Telephone and Telegraph Co. from 1905 to 1926. (Data, Table 27).

in each year. The band or zone between these curves is colored. This band gives a good graphic picture of the ups and downs of the stock over a series of years. This is illustrated in fig. 21, plotted from the data in table 27.

**47. Cumulative curves.**—Let us consider the following set of data on number of farms by size in 1920.

**Table 28.—Number of Farms in the United States in 1920 by Size.**

| *Size in Acres* | *Number* | *Size in Acres* | *Number* |
|---|---|---|---|
| Under 10 | 288,772 | 175 to 259 | 530,800 |
| 10 to 19 | 507,763 | 260 to 499 | 475,677 |
| 20 to 49 | 1,503,732 | 500 to 999 | 149,819 |
| 50 to 99 | 1,474,745 | 1,000 and over | 67,405 |
| 100 to 174 | 1,449,630 | Total | 6,448,343 |

*Statistical Abstract of U. S., 1924, p. 564.*

We construct from this table another in which is given the number of farms having "less than" a stated number of acres. In plotting, the frequencies are plotted at the upper limits of the class. Each count gives the number of farms (items) in the given class and all lower classes. The data is said to be cumulated on the "less than" basis. This is illustrated by table 29.

We construct from this table still another in which is given the number of farms having "more than" a stated number of acres. In plotting, the frequencies are plotted at the lower limits of the class. Each count gives the number of farms (items) in the given class and all higher classes. The data are said to be cumulated on the "more than" basis. This is illustrated by table 30.

In both cases we are said to have a cumulative frequency distribution.

Let us plot the data given in table 29 and table 30. Use acreage as abscissas and number of farms as ordinates. Two $S$-shaped curves are formed as shown in fig. 22. The $S$-shape of the curves in fig. 22 is characteristic of all cumulative curves.

**Tables 29–30.—Number of Farms in the United States in 1920 by Size.**

| *Acreage less than* | *Number* |
|---:|---:|
| 10 | 288,772 |
| 20 | 796,535 |
| 50 | 2,300,267 |
| 100 | 3,775,012 |
| 175 | 5,224,642 |
| 260 | 5,755,442 |
| 500 | 6,231,119 |
| 1,000 | 6,380,938 |
| Total............ | 6,448,343 |

**Table 29.—"Less Than" Cumulation.**

| *Acreage more than* | *Number* |
|---:|---:|
| 999 | 67,405 |
| 499 | 217,224 |
| 259 | 692,901 |
| 174 | 1,223,701 |
| 99 | 2,673,331 |
| 49 | 4,148,076 |
| 19 | 5,651,808 |
| 9 | 6,159,571 |
| Total............ | 6,448,343 |

**Table 30.—"More Than" Cumulation**

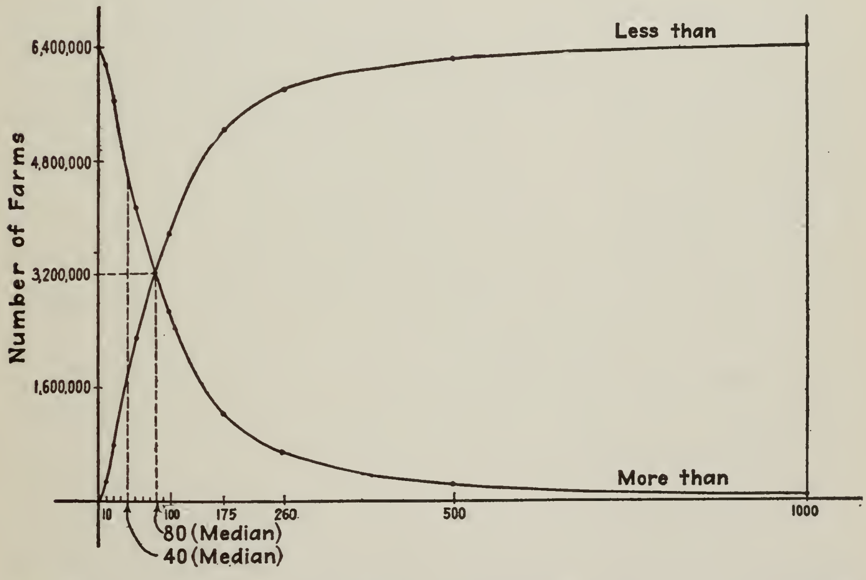


**Fig. 22.—Cumulative Graph Showing the Number of Farms in the United States, in 1920, Less Than and More Than a Given Size.**

This type of curve is useful in finding the median and the mode.[4] The median can be determined graphically as the point of intersection of the two cumulative graphs. When

[4] See Chapter VII for definitions of median and mode.

only one graph is given, the median can be determined as that abscissa corresponding to that ordinate which bisects the total frequency. In this case the mid-ordinate is 3,224,172 or half the total number of farms. The corresponding abscissa, as read from the graph, is approximately 80 (computation gives 81+). Thus, from the graph we would conclude that there are just as many farms containing less than 80 acres as there are farms which contain more than 80 acres.

The modal size of farm (that size of farm most frequently occurring) is determined from the graph by determining the abscissa of the "point of inflection" on the curve. A point of inflection is a point at which a tangent line, in sliding along the curve, quits turning in one direction and starts to turn in the opposite direction. For the curve in fig. 22 the mode is determined at approximately 40 acres.

The curve is useful also in comparing two or more sets of data having irregular or dissimilar class intervals. The irregularity of the intervals has little effect upon the final curve. In the present instance, whatever the size of the intervals, the total or cumulative frequency for farms less than 260 acres must remain the same.

The cumulative frequency distribution and curve find practical application in keeping track of the total output of factories both as to quantity and value. The cumulative curve is used in tabulating the number of employees receiving "less than" or "more than" a given wage; the number of people receiving "less than" or "more than" a given income; the number of employees who have been in the employ of a stated firm "more than" or "less than" a given time; the number of families with "more than" or "less than" a given number of children; the total sales of a factory or wholesale firm to date, or the total sales in a given territory.

### Exercises.

Plot cumulative frequency graphs for the following data (*source: Statistical Abstract of the U. S., 1924*):

1 *p. 760.* Average number of wage earners, distributed according to prevailing hours of labor per week, for major industries in 1923.

| NDUSTRY | Hours of Labor per Week | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 44 and under | 44<br>48 | 48 | 48<br>54 | 54 | 54<br>60 | 60 and over | Total |
| *Boots and Shoes* | 19,724 | 19,119 | 100,314 | 61,588 | 16,864 | 7,484 | 123 | 225,216 |
| *Boxes, Paper and Other, not Otherwise Specified* | 3,226 | 3,570 | 19,565 | 21,837 | 2,764 | 5,288 | 605 | 56,855 |
| *Confectionery and Ice-cream* | 3,224 | 5,077 | 18,873 | 19,576 | 8,240 | 6,577 | 1,918 | 63,485 |
| *Electrical Machinery, Apparatus, and Supplies* | 15,187 | 8,256 | 125,806 | 65,142 | 8,622 | 11,508 | 371 | 234,892 |
| *Glass* | 8,150 | 8,663 | 37,393 | 7,814 | 4,650 | 4,599 | 2,066 | 73,335 |
| *Stoves and Hot-air Furnaces* | 2,526 | 963 | 13,327 | 7,747 | 5,272 | 1,959 | 1,235 | 33,029 |

2. *p. 788.* Number of building contracts awarded by months in 27 Northern and Eastern states.

| *Months* | Number of Projects | | | *Months* | Numbfr of Projects | | |
|---|---|---|---|---|---|---|---|
| | *1922* | *1923* | *1924* | | *1922* | *1923* | *1924* |
| January | 5,073 | 6,126 | 6,752 | July | 9,902 | 7,925 | 8,556 |
| February | 4,782 | 6,338 | 6,571 | August | 10,457 | 8,381 | 9,013 |
| March | 9,250 | 10,546 | 9,986 | September | 9,108 | 7,500 | 9,035 |
| April | 10,746 | 12,336 | 11,021 | October | 9,568 | 9,844 | 9,981 |
| May | 11,358 | 11,536 | 10,938 | November | 9,079 | 8,794 | 9,219 |
| June | 11,249 | 8,372 | 9,454 | December | 7,080 | 7,757 | 8,526 |

3. *p. 565.* Per cent distribution of farm land by size of farm.

| | Improved | | | All | | |
|---|---|---|---|---|---|---|
| | *1900* | *1910* | *1920* | *1900* | *1910* | *1920* |
| Under 20 acres | 1.6 | 1.7 | 1.6 | 0.9 | 1.0 | 0.9 |
| 20 to 49 acres | 8.0 | 7.6 | 7.7 | 5.0 | 5.2 | 5.1 |
| 50 to 99 acres | 16.2 | 14.9 | 14.4 | 11.8 | 11.7 | 11.1 |
| 100 to 174 acres | 28.6 | 26.9 | 25.5 | 23.0 | 23.4 | 20.4 |
| 175 to 499 acres | 32.7 | 33.8 | 33.8 | 27.8 | 30.2 | 29.0 |
| 500 to 999 acres | 7.1 | 8.5 | 9.6 | 8.1 | 9.5 | 10.6 |
| 1,000 acres and over | 5.9 | 6.5 | 7.5 | 23.6 | 19.0 | 23.1 |

4. *p. 378.* Car loadings: Average weekly loadings (by months) of revenue freight, Class I Railways, by districts and by principal commodity groups for the years 1920–1924 inclusive.

5. *p. 368.* Railroad revenue and traffic by months for class I carriers, for the years 1910–1924 inclusive.

6. *p. 70.* Deaths of infants under 1 year of age. Rate per 1,000 births according to age subdivisions, for the years 1917–1923 inclusive.

7. *p. 62.* Death rates per 1,000 population by sex and by age groups for the years 1900, 1910, 1920–1923 inclusive.

8. *p. 46.* Persons engaged in gainful occupations, by sex and by age for 1910 and 1920; by sex, by age, and principal classes for 1920.

9. *p. 13.* Age distribution of total population: by classes, 1920, with certain comparisons for previous censuses.

10. *p. 294.* Commercial failures: (a) Number by months for a given year. Years tabulated are 1918–1924 inclusive. (b) Liabilities (1,000 dollars) by months for a given year.

**48. Lorenz curves.**—Let us consider the following data:

**Table 31.—Manufactures: Size of Establishment as Measured by the Value of Products for 1921.**

| (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|
| Value of Products per Establishment | Establishments | | Value of Products | |
| | *Number* | *Per Cent of Total* | *Dollars (Millions)* | *Per Cent of Total* |
| Less than $5,000........ | 53,999 | 21.6 | 136.9 | 0.3 |
| 5,000– 20,000........ | 71,075 | 28.4 | 783.0 | 1.8 |
| 20,000– 100,000........ | 72,251 | 28.9 | 3,330.3 | 7.6 |
| 100,000–1,000,000........ | 45,608 | 18.2 | 13,702.4 | 31.3 |
| 1,000,000 and over.......... | 7,333 | 2.9 | 25,837.4 | 59.0 |
| Total..................... | 250,266 | 100.0 | 43,790.0 | 100.0 |

*Statistical Abstract of the U. S., 1924, p. 724.*

Let us form a table of the cumulated frequencies.

**Table 32.—Cumulative Data Showing Size of Establishment as Measured by Value of Products for 1921 (see Data, Table 31).**

| *a* | *b* | *c* |
|---|---|---|
| Value of Products per Establishment | Establishments | Value of Products |
| | *Per Cent* | *Per Cent* |
| Less than $5,000.............. | 21.6 | 0.3 |
| " " 20,000.............. | 50.0 | 2.1 |
| " " 100,000.............. | 78.9 | 9.7 |
| " " 1,000,000.............. | 97.1 | 41.0 |
| Any value whatever.............. | 100.0 | 100.0 |

One can make a frequency graph with the data in columns (1) and (3) in table 31, and another frequency graph with the data in columns (1) and (5) in the same table.

One can make a cumulative frequency graph with the data in columns (a) and (b) in table 32. One can make another cumulative frequency graph with the data in columns (a) and (c) in the same table. These cumulative graphs have been considered in the previous article.

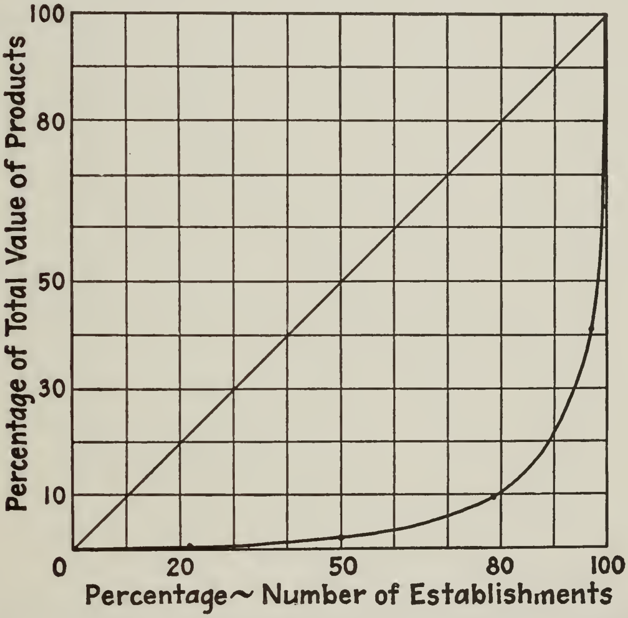


Fig. 23.—Lorenz Curve, Showing Percentage of Number of Establishments Manufacturing Percentage of Total Value of Products. Data, Table 32.

A new graph can be made by using the data in columns (b) and (c) in table 32 as abscissas and ordinates, respectively. This graph is known as a Lorenz curve in honor of Dr. Max O. Lorenz, who first devised it.

The chart omits entirely the classes which have been used to group the frequencies. All items are turned into percentages of the total and plotted as percentages on both axes.

The graph shows that 50 per cent of the total number of establishments do not produce 50 per cent of the total value of products.

Curves of this type are used to show distribution of wealth, income, rents, and wages.

Table 33, which is taken from King's "Wealth and Income of the People of the United States," p. 71, gives the cumulative percentages of the number and value of estates of men dying in Massachusetts for the period 1889 to 1891.

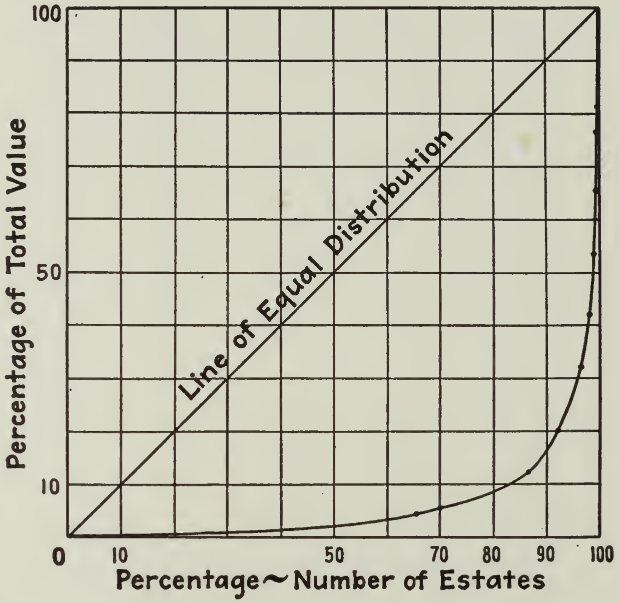


Fig. 24.—Lorenz Curve Showing Cumulative Percentage of the Number and Value of Estates of Men Dying in Massachusetts in 1889–1891. Data, Table 33.

If 10 per cent of the estates had 10 per cent of the wealth, 25 per cent had 25 per cent, 50 per cent had 50 per cent, and 80 per cent had 80 per cent, and so on, then wealth would be evenly distributed. The corresponding graph would be the straight line marked in fig. 24 as the line of equal distribution. The graph (and data) show that 70 per cent of the people possess but 5 per cent of the wealth, while the richest 1 per cent have about 47 per cent of the total wealth. The farther the curve departs from the straight line, the more unequal the distribution.

**Table 33.—Cumulative Percentage of the Number and Value of Estates of Men Dying in Massachusetts, 1889–1891.**

| Value of Estate in Thousands of Dollars | Estates of Given Value | |
|---|---|---|
| | *Per Cent of Total Number* | *Per Cent of Total Value* |
| Value 0.5 | 65.864 | 4.568 |
| 1 | 70.036 | 5.248 |
| 5 | 86.298 | 12.862 |
| 10 | 92.002 | 20.082 |
| 25 | 96.588 | 32.087 |
| 50 | 98.116 | 42.131 |
| 100 | 99.066 | 53.716 |
| 200 | 99.567 | 65.629 |
| 300 | 99.735 | 72.847 |
| 400 | 99.815 | 77.530 |
| 500 | 99.866 | 81.610 |
| All estates | 100.000 | 100.000 |

## Exercises.

1. Plot Lorenz curves for the following data (*source: King's "Wealth and Income"*):

(a) *p. 88.* Estates probated in France in 1909.

| Value of Estate in Francs | Estates of Given Value | |
|---|---|---|
| | *Per Cent of Total Number* | *Per Cent of Total Value* |
| *Excess of debts* | 3.533 | |
| 500 or less | 29.834 | 0.470 |
| 2,000 | 55.556 | 2.734 |
| 10,000 | 83.632 | 12.198 |
| 50,000 | 96.028 | 30.079 |
| 100,000 | 97.984 | 39.303 |
| 250,000 | 99.210 | 52.521 |
| 500,000 | 99.647 | 63.072 |
| 1,000,000 | 99.853 | 72.730 |
| 2,000,000 | 99.948 | 81.652 |
| 5,000,000 | 99.985 | 89.066 |
| 10,000,000 | 99.997 | 94.350 |
| 50,000,000 | 99.997 | 97.485 |
| All estates | 100.000 | 100.000 |

(b) *p. 92.* Prussian families' wealth from tax assessments for 1908.

| Family Wealth in Marks | Estates of Given Value | |
|---|---|---|
| | *Per Cent of Total Number* | *Per Cent of Total Value* |
| Less than 6,000 | 85.880 | 13.76 |
| 20,000 | 92.684 | 22.02 |
| 52,000 | 97.581 | 38.53 |
| 100,000 | 98.901 | 49.54 |
| 200,000 | 99.509 | 59.63 |
| 500,000 | 99.839 | 71.56 |
| 1,000,000 | 99.939 | 79.82 |
| 5,000,000 | 99.995 | 92.66 |
| All families | 100.000 | 100.00 |

(c) *p. 77.* Cumulative percentages of the number and value of estates of men dying in the year 1900 in six Wisconsin counties.

(d) *p. 90.* Distribution of wealth among men over twenty-five years of age dying in the United Kingdom, 1907–1911.

(e) *p. 233.* Distribution of income among Prussian population in 1910.

(f) *p. 224.* Distribution of income among the families of the Continental United States in 1910.

2. *Jerome, Harry, "Statistical Method," Harper's, 1924, p. 150.* Distribution of taxable income and of the tax thereon among the persons assessed to Wisconsin income tax on income of the year 1919.

3. *Day, E. E., "Statistical Analysis," Macmillan, 1925, p. 88.* Data for Lorenz curve of wage earners and wage receipts in given wage earning group.

4. For distribution of labor incomes of farmers, see *"Income in the United States" Vol. II of the series published by the National Research Council.*

5. *Statistical Abstract of the U. S., 1924, p. 724.* Data for the years 1904, 1909, 1914, 1919, 1923, on size of manufacturing establishment as measured by value of products.

6. *Statistical Abstract of the U. S., 1924, p. 162.* (a) Number of returns and net income. (b) Number of returns and tax yield.

**49. Logarithmic charts.**—These charts are sometimes known as *ratio* charts.

Let us consider the following data on the wholesale New York average price in cents of fresh milk per quart and creamery butter, extra, per lb., 1913–1925 inclusive, as given in Bulletin No. 415, Bureau of Labor Statistics (milk, p. 60; butter, p. 78).

**Table 34.—Price of Milk and Butter, 1913–1925.**

| *Year*....... | *1913* | *14* | *15* | *16* | *17* | *18* | *19* | *1920* | *21* | *22* | *23* | *24* | *25* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Milk....... | 4.4 | 4.2 | 4.2 | 4.5 | 6.3 | 7.8 | 8.4 | 8.5 | 7.5 | 7.3 | 7.6 | 6.8 | 7.6 |
| Butter..... | 32.3 | 29.9 | 29.9 | 34.1 | 42.7 | 51.6 | 60.5 | 61.4 | 43.4 | 40.6 | 46.8 | 42.7 | 45.4 |

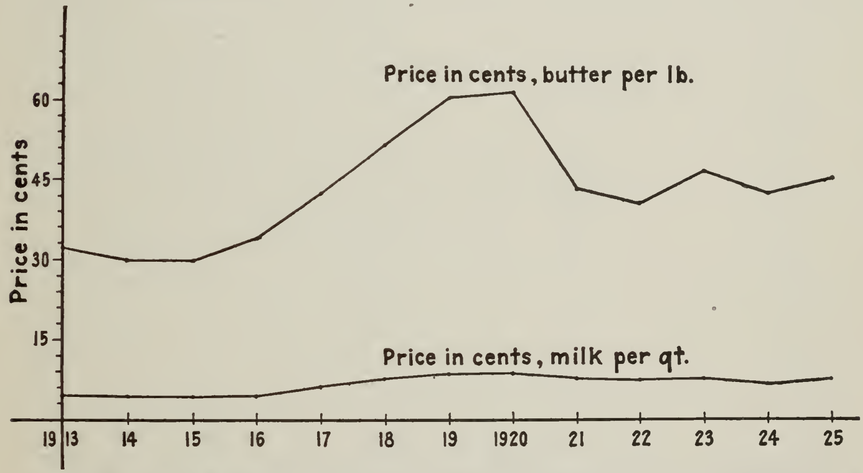


**Fig. 25.—Graph Showing Wholesale New York Average Price in Cents of Fresh Milk per Quart and Creamery Butter, Extra, per Pound, 1913–1925. Data, Table 34.**

Fig. 25 is made on the basis of the actual changes in price. From the graph it would appear that the changes in the price of butter had been more violent than the changes in the price of milk. The greatest percentage change in the price of butter is 205 from 1915 to 1920. The greatest percentage change in milk is 202 and is for the same years, 1915 to 1920.

What is needed is a chart which will represent equal percentage changes by equal vertical changes in the ordinates. This is accomplished by means of a logarithmic chart.

The *difference* between any two consecutive numbers of the following list:

(a) 0 1 2 3 4 5

is a constant, namely unity. These are plotted on the natural scale in fig. 26. The *ratio* between any two consecutive numbers of the following list:

| (b) | 1 | 10 | 100 | 1,000 | 10,000 | 100,000 |
|---|---|---|---|---|---|---|
| | $10^0$ | $10^1$ | $10^2$ | $10^3$ | $10^4$ | $10^5$ |

is a constant, namely 10. If we express each number, as shown, as a power of 10, the exponents are the first set of numbers shown and these exponents have a constant difference. These exponents are the logarithms, to the base 10, of the corresponding numbers. Thus, if $10^2 = 100$, we say that the logarithm of 100 to the base 10 is 2.

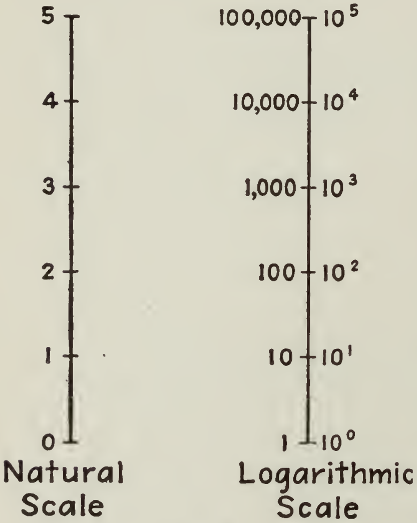


Fig. 26.

In order to plot on a logarithmic scale the series of numbers (b), find the logarithm to the base ten, of each number. In this case we obtain the set of numbers (a). Plot the series (a) on a natural scale as shown in fig. 26. Instead, however, of placing the numbers 0, 1, 2, 3, 4, 5 at the equally spaced intervals, put the set of numbers (b), equally spaced, as shown in fig. 26 on the vertical line labeled "logarithmic scale." As shown in fig. 26, the numbers which are equally spaced have a common ratio. The scale, in practice, is labeled with numbers whose absolute differences are equal, and hence the numbers are not equally spaced. Thus, the natural numbers, 10, 20, 30, 40, 50, 60, 70, 80, whose logarithms are 1.00, 1.30, 1.47, 1.60, 1.70, 1.78, 1.85, 1.90 would be spaced as shown in fig. 27.

The vertical scale in fig. 28, as shown on the left in column (a), has been constructed after the manner indicated in fig. 27. The columns of figures (b) and (c) illustrate how different scales may be placed on the same chart. The figure also illustrates the periodic manner of ruling a logarithmic chart. Each period ends with a power of ten. The sample in fig. 28 is called a semi-logarithmic chart, for the scale in only one direction is logarithmic. The scale in the other direction is the natural scale. When the scale is logarithmic in both directions, the chart is said to be double logarithmic.

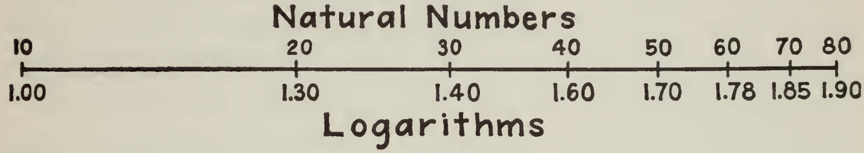


Fig. 27.

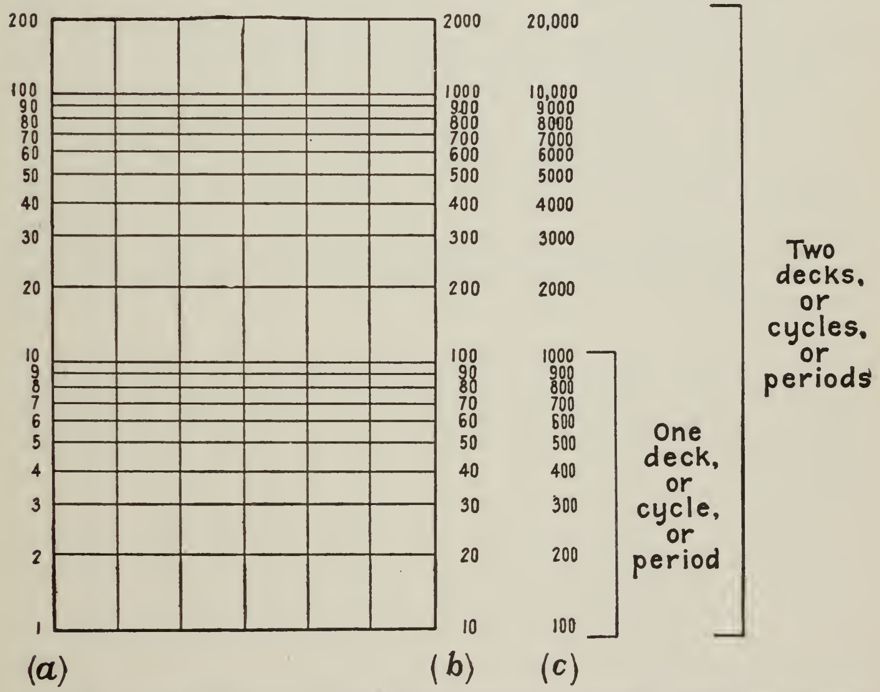


Fig. 28.—Graphs Illustrating Rulings on Semi-logarithmic Coördinate Paper.

Let us plot on a semi-logarithmic chart, fig. 29, the data in table 34 on butter per lb. and milk per quart.

On a logarithmic chart these two curves are nearly parallel throughout. This suggests that the changes in one have been, in general, the same as the changes in the other. This represents the facts in the case when we think of percentage changes.

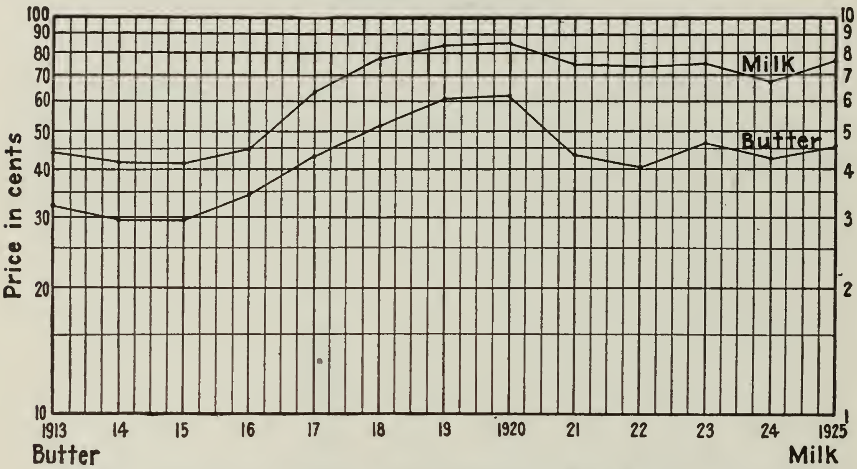


Fig. 29.—Semi-logarithmic Graph of Wholesale New York Average Price in Cents of Fresh Milk per Quart and Creamery Butter, per Pound, 1913–1925. Data, Table 34.

Let us now plot on the natural scale and again on a semi-logarithmic chart the growth of $1 and $5 at compound interest at 6 per cent for 0, 10, 20, 30, 40, 50 years. (See fig. 30.)

| *Years* | *0* | *10* | *20* | *30* | *40* | *50* | *60* |
|---|---|---|---|---|---|---|---|
| $1 | 1 | 1.791 | 3.207 | 5.743 | 10.286 | 18.420 | 32.987 |
| $5 | 5 | 8.95 | 16.03 | 28.71 | 51.43 | 92.10 | 164.83 |

The curves plotted on the natural scale do not suggest equal rates of growth. The same data plotted on a logarithmic scale give two parallel lines. This suggests equal rates of growth. This comes from the fact that equal vertical intervals on a logarithmic scale represent equal ratios.

Whenever rates are of paramount interest, a logarithmic scale is useful. The chief advantages of a semi-logarithmic chart may be summarized as follows:

1. If a curve is ascending and is nearly straight, then the magnitude which it represents is increasing at an almost constant rate.

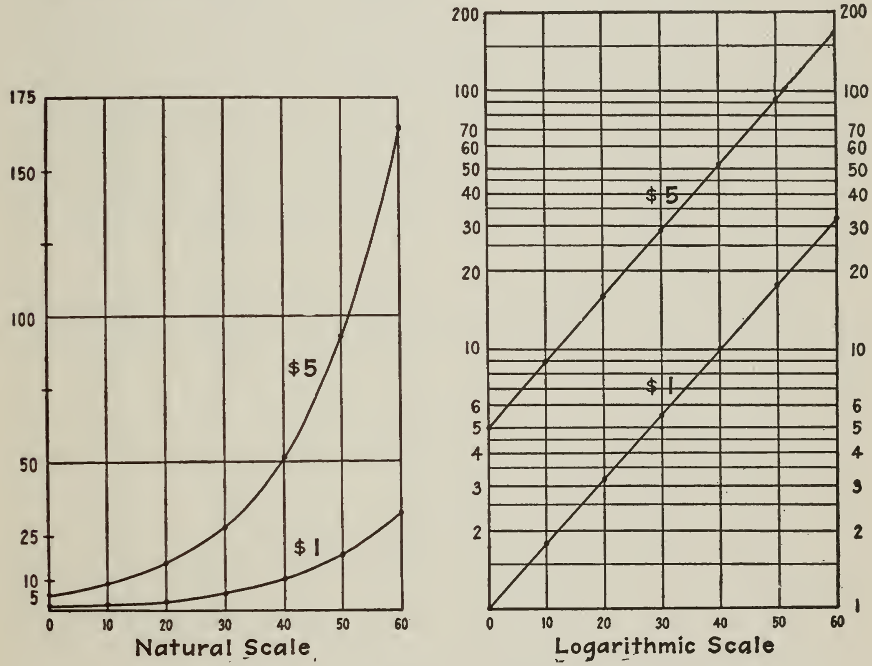


**Fig. 30.—$1 and $5 at 6% Compound Interest.**

2. If a curve is descending and is nearly straight, then the magnitude which it represents is decreasing at an almost constant rate.

3. If a curve bends upward, the rate of growth is increasing.

4. If a curve bends downward, the rate of growth is decreasing.

5. If two curves are parallel, the rate of growth is the same for both.

6. Of two curves, the steeper represents the greater rate of growth.

7. Of two parts of the same curve, the steeper part represents the greater rate of growth.

A semi-logarithmic chart is useful in plotting data with an extreme range in one direction. If the range is extreme in both directions, a double logarithmic chart is useful.

Negative numbers cannot be plotted on a logarithmic chart, for the logarithm of a negative number is not a real number. For some data negative numbers are not possible; for example, population, production of a factory, sales. Temperature has an arbitrary zero. Any temperature can be plotted by using absolute zero as a base. For time, the zero is arbitrary and hence any time can be plotted on a logarithmic scale. Zero for the original data cannot be shown, for $\log 0 = -\infty$

Logarithmic charts are an aid in detecting correlation and in forecasting, especially when the curve on a logarithmic chart is a straight line.

**Exercises.**

Plot logarithmic curves for the following data (*source: U. S. Bureau of Labor Statistics, Bulletin No. 334*):

1. *p. 198.* Average retail price of sugar in cents per lb., New York.

*p. 206.* Average retail price of creamery butter in cents per lb., New York.

| Year | *1913* | *14* | *15* | *16* | *17* | *18* | *19* | *20* | *21* | *22* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Sugar*.................... | 4.9 | 4.7 | 5.2 | 6.3 | 7.4 | 9.7 | 10.1 | 17.3 | 9.0 | 5.2 |
| *Butter*.................... | 36.2 | 38.1 | 36.3 | 36.7 | 43.8 | 54.4 | 71.3 | 69.0 | 56.3 | 40.8 |

2. *p. 728 of Yearbook of U. S. Department of Agriculture, 1925.* Farm price per bu. for corn, Dec. 1.

*p. 1128.* Farm price per 100 pounds for hogs, Dec. 1.

| Year | *1910* | *11* | *12* | *13* | *14* | *15* | *16* | *17* | *18* | *19* | *20* | *21* | *22* | *23* | *24* | *25* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Corn* $ | .48 | .618 | .487 | .691 | .644 | .575 | .889 | 1.279 | 1.365 | 1.345 | .670 | .423 | .658 | .726 | .982 | .674 |
| *Hogs* $ | 7.16 | 5.72 | 6.89 | 7.16 | 6.67 | 6.02 | 8.76 | 15.73 | 15.82 | 12.66 | 8.90 | 6.52 | 7.63 | 6.39 | 8.39 | 10.51 |

3. *p. 262 of Statistical Abstract of the U. S., 1924.* New York clearing house transactions in billions of dollars.

*p. 743 of Yearbook of U. S. Department of Agriculture, 1925.* Wheat produced in millions of bushels. For years 1900–1908 inclusive on wheat, see *ibid.* for 1909.

| Year | *N. Y. Clearings* | *Wheat* |
|---|---|---|
| 1900 | 52.0 | 522 |
| 01 | 77.0 | 748 |
| 02 | 74.8 | 670 |
| 03 | 70.8 | 638 |
| 04 | 59.7 | 552 |
| 1905 | 91.9 | 693 |
| 06 | 103.8 | 735 |
| 07 | 95.3 | 634 |
| 08 | 73.6 | 665 |
| 09 | 99.3 | 700 |
| 1910 | 102.6 | 635 |
| 11 | 92.4 | 621 |
| 12 | 96.7 | 730 |
| 1913 | 98.1 | 763 |
| 14 | 89.8 | 891 |
| 1915 | 90.8 | 1,026 |
| 16 | 147.2 | 636 |
| 17 | 181.5 | 637 |
| 18 | 174.5 | 921 |
| 19 | 214.7 | 968 |
| 1920 | 252.3 | 833 |
| 21 | 204.1 | 815 |
| 22 | 213.3 | 868 |
| 23 | 214.6 | 797 |
| 24 | 235.5 | 863 |

4. *pp. 156, 158, 160 of U. S. Bureau of Labor Statistics, Bulletin No. 415, 1890–1925.* Relative prices of white oak, plain; yellow pine siding: hemlock No. 1, northern.

5. *p. 82, U. S. Bureau of Labor Statistics, Bulletin No. 415, 1890–1925.* Loaf bread, New York, relative price.

*p. 88.* Wheat flour, New York spring patents and Minneapolis standard patents, relative price.

6. *p. 72, Statistical Abstract of the U. S., 1924.* Marriages and divorces, total number.

*pp. 2 and 3.* Population.

7. *Source: Statistical Abstract of the U. S., 1924.*

(a) *p. 783.* Production of manufactured tobacco.

(1) Cigarettes, small.
(2) Plug tobacco.
(3) Cigars; (a) large; (b) small.
(4) Twist.
(5) Fine cut.

(b) *p. 380.* Persons killed and injured in railroad accidents.

(1) Passengers killed, employees killed, on same chart.
(2) Passengers injured, employees injured, on same chart.

(c) *p. 696.* Production of pig iron in the United States, 1810–1900. Production of bituminous coal, *p. 704.*

(d) *p. 350.* Production and registration of motor vehicles.

(e) Use double logarithmic paper to chart number of persons reporting income in *excess* of specified amounts. *Data table XIV, appendix.*

**50. Pareto's law.**—An asymmetrical curve with a long tail to the right tends to become symmetrical when plotted on semi-logarithmic paper, plotting log $X$ in place of $X$.

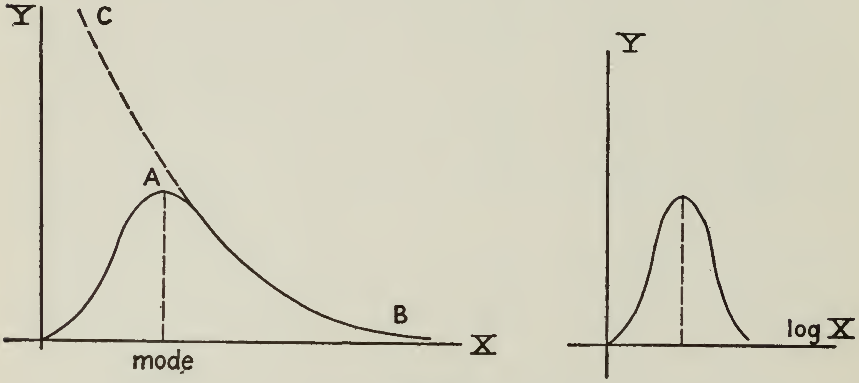


Fig. 31.

Thus, in plotting $y$ = number of farms, $x$ = number of acres (data table 28) the range for $x$ is from 3 acres to over 1,000 acres, and there is a long tail to the right. Data with a large range in one direction are common in business and economic statistics.

Plot that part of the data which is represented by that part of the curve ($AB$) to the right of the mode,[1] using double logarithmic paper. The resulting curve tends to be a straight line. This tendency was first noticed by Vilfredo Pareto, an Italian economist, in connection with the distribution of income. The portion $AB$ of the curve, if extended to the left along the dotted line $AC$, can be approximated by the curve whose equation is

$$y = Cx^{-m}$$

[1] See Chapter VII for definition of mode.

Taking logarithms, we find:

$$\log y = -m \log x + \log C$$

If we put:

$$Y = \log y; X = \log x; B = \log C$$

we have:

$$Y = -mX + B$$

which is the equation of a straight line with negative slope.

The National Bureau of Economic Research has made an extensive study of the distribution of income and has examined Pareto's law in detail. The results of this study have been published in two volumes: "Income in the United States." It would seem that Pareto's conclusions must be considered as a rough first approximation.

## Chapter VII

# AVERAGES

**51. Introduction.**—A set of data, if the items are at all numerous, contains so much detail that its characteristics are not easily seen. Some summarizing expression is needed, some single term, which by itself conveys an adequate idea of the group for which it stands. Several such summarizing expressions have come into common use. Of these we shall discuss, in this chapter, averages; in subsequent chapters, dispersion, and skewness.

We shall discuss the following averages:

1. Mid-range value.
2. Arithmetic average.
3. Mode.
4. Median.
5. Geometric average.
6. Harmonic mean.
7. Root mean square.

For each of these averages, we shall give its (a) definition; (b) method of computation; and (c) merits and weaknesses.

**52. Mid-range.**—In constructing table 35, all items which measured 4 feet or more, and less than 5 feet, were counted and listed, as the frequency 7, opposite the class interval "4 and under 5." It is possible that no two of these 7 items had the same length. In using this table, in the absence of the original data, one assumes that the 7 distinct items of length varying from 4 to 5 feet can be thought of as 7 items each of the same length, namely 4.5 feet, the mid-point of the interval. In the same way, one can think of the class interval from 3 feet to 10 feet with a frequency of 217 as the equivalent of 217 items each of the same length, namely, the length corresoonding to the mid-point of the interval, 6.5 feet.

Table 35.—Hypothetical Frequency Distribution of 217 Items.

| *h in Feet* | *Number of Items* | *Cumulative Frequency* |
|---|---|---|
| 3 and under 4 | 3 | 3 |
| 4 " " 5 | 7 | 10 |
| 5 " " 6 | 22 | 32 |
| 6 " " 7 | 60 | 92 |
| 7 " " 8 | 85 | 177 |
| 8 " " 9 | 32 | 209 |
| 9 " " 10 | 8 | 217 |
| | 217 | |

In general, if $l$ = lowest limit of the range, and $L$ = upper limit of the range,

$$\text{Mid-range} = \frac{l + L}{2}$$

This average is little used and in general is not a good representative number. It depends for its value upon but two numbers, namely the limits of the range under consideration, and takes no account of the distribution within that range. It is easily computed.

**53. Arithmetic average.**—This is the most frequently used average. It is the result obtained when the sum of the measurements of the items in a group is divided by the number of items in the group. This average is at times referred to as the *arithmetic mean* or simply the *mean*.

If we let $x$ represent the variable measurement of the individual items, $x_i$ represent the measurement of the $i$-th item, $N$ the number of items, $\overline{X}$ the arithmetic average, then from the definition of the arithmetic average, we have

$$\overline{X} = \frac{x_1 + x_2 + \cdot \cdot \cdot + x_n}{N} = \frac{\Sigma x^*}{N}$$

* $\Sigma x$ stands for the sum of all of the values of $x$ which represent the measures of the individual items. $\Sigma$ is the Greek *sigma* and stands for summation.

If we are computing the arithmetic average from a frequency distribution, in the absence of the original data, we consider that all of the items enumerated in a given class interval have the same measurement, namely the measurement represented by the mid-point of the class interval, or what amounts to the same thing, that the arithmetic average of the frequencies within the interval is the mid-point of the interval. The sum of the measurements of the items in any class interval is then the product of the frequency and the measure of the mid-point of that interval. Compute this product for each class interval, add the results and divide by the total frequency.

If $x_i$ represents the mid-measure of the $i$-th class interval, $f_i$ the corresponding frequency, and $n$ the number of class intervals, then

$$\overline{X} = \frac{\sum_1^n f_i x_i}{\sum_1^n f_i}$$

In order not to complicate the notation, we frequently write more simply

$$\overline{X} = \frac{\Sigma fx}{\Sigma f} = \frac{\Sigma fx}{N}$$

For the hypothetical distribution given in the previous article we have the computations shown in table 36.

**Table 36.—Computation of Arithmetic Average.**

| *Mid-point of Interval* $x_i$ | *Frequency* $f_i$ | $f_i x_i$ | |
|---|---|---|---|
| 3.5 | 3 | 10.5 | $\overline{X} = \frac{1538.5}{217}$<br>$= 7.09$ |
| 4.5 | 7 | 31.5 | |
| 5.5 | 22 | 121.0 | |
| 6.5 | 60 | 390.0 | |
| 7.5 | 85 | 637.5 | |
| 8.5 | 32 | 272.0 | |
| 9.5 | 8 | 76.0 | |
| | 217 | 1,538.5 | |

**54. Short-cut method.**—In many instances, it is convenient to use a short-cut process. The steps are as follows:

1. *Assume an average.* By an inspection of the frequency table, one can guess approximately the average. Select some convenient number near this approximation. In general, select the mid-point of some class interval as an approximate average.

2. Find the deviations ($\xi$) from this assumed average of the mid-points of the several class intervals.

3. Multiply each class frequency ($f$) by the deviation ($\xi$) of the mid-point of its class interval from the assumed average.

4. Add these products ($\Sigma f\xi$) and divide by the total frequency ($N$).

5. Add this quotient to the assumed average $x_o$.

To prove that this is the correct procedure, let us call the assumed average $x_o$. Let $\xi_i$ be the deviation of the mid-point of the $i$-th class interval from the assumed average. Then

$$x_i = x_o + \xi_i$$

$$\Sigma f_i x_i = \Sigma f_i x_o + \Sigma f_i \xi_i$$

$$\frac{\Sigma f_i x_i}{N} = \frac{\Sigma f_i x_o}{N} + \frac{\Sigma f_i \xi_i}{N}$$

$$\therefore \overline{X} = x_o + \frac{\Sigma f_i(x_i - x_o)}{N}$$

since

$$\frac{\Sigma f_i x_i}{N} = \overline{X}; \frac{\Sigma f_i x_o}{N} = x_o \frac{\Sigma f_i}{N} = x_o; \xi_i = x_i - x_0$$

This formula is sometimes more simply written as

$$\overline{X} = x_o + \frac{\Sigma f(x - x_o)}{N}$$

Applying this to the hypothetical distribution in table 35, assume $x_o = 6.5$. The computations are systematically tabulated, as shown in table 37.

**Table 37.—Computation of Arithmetic Average.**

| *Interval*<br>$x_i$ | *Frequency*<br>$f_i$ | $x_i - x_o$ | $f_i(x_i - x_o)$ |
|---|---|---|---|
| 3.5 | 3 | −3 | − 9 |
| 4.5 | 7 | −2 | −14 |
| 5.5 | 22 | −1 | −22 |
| 6.5 | 60 | 0 | 0 |
| 7.5 | 85 | 1 | 85 |
| 8.5 | 32 | 2 | 64 |
| 9.5 | 8 | 3 | 24 |
| | $N = 217$ | | 128 |

$\overline{X} = 6.5 + \frac{128}{217} = 6.5 + 0.59 = 7.09$

**55. Step deviation.**—Whatever the class interval, call it unity. Compute the correction to the assumed average by the ordinary means and then multiply by the class interval. If we let

$i$ = class interval

$s$ = step deviation from the assumed mean

then

$$\overline{X} = x_o + \frac{\Sigma fs}{N} i$$

This method will be illustrated by computing the average egg production from the frequency distribution in table 38 on first year egg production as given in Bulletin 110, Part 1, Bureau of Animal Husbandry, United States Department of Agriculture.

Table 38.—Frequency Table of First Year Egg Production in the Domestic Fowl.

| *Egg Production Class Interval* | *No. of Fowls Frequency* $f_i$ | *Mid-point of Interval* $x_i$ | *Step Deviation* $s_i$ | $f_i s_i$ |
|---|---|---|---|---|
| 0– 14 | 0 | 7 | −8 | 0 |
| 15– 29 | 2 | 22 | −7 | −14 |
| 30– 44 | 0 | 37 | −6 | 0 |
| 45– 59 | 1 | 52 | −5 | − 5 |
| 60– 74 | 5 | 67 | −4 | −20 |
| 75– 89 | 8 | 82 | −3 | −24 |
| 90–104 | 17 | 97 | −2 | −34 |
| 105–119 | 18 | 112 | −1 | −18 |
| 120–134 | 17 | 127 | 0 | 0 |
| 135–149 | 26 | 142 | 1 | +26 |
| 150–164 | 17 | 157 | 2 | 34 |
| 165–179 | 18 | 172 | 3 | 54 |
| 180–194 | 9 | 187 | 4 | 36 |
| 195–209 | 2 | 202 | 5 | 10 |
| 210–224 | 6 | 217 | 6 | 36 |
| 225–239 | 1 | 232 | 7 | 7 |
| | $N = 147$ | | | $\Sigma fs = 88$ |

$$\overline{X} = 127 + 15 \times \tfrac{88}{147} = 127 + 8.98 = 135.98; \; x_0 = 127$$

**56. Properties of an arithmetic average.**—Each of the following series

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | 7 | 7 | 7 | 7 | | |
| 5 | 6 | 7 | 8 | 9 | | |
| 4 | 5 | 6 | 8 | 9 | 10 | |
| 1 | 1 | 2 | 2 | 3 | 5 | 35 |
| 2 | 12 | | | | | |

has 7 as an arithmetic average. Thus we see that an arithmetic average tells us nothing with respect to the distribution of the items. This average is independent of the order of the items, the number of items, and relative size of the items. Each item is of equal importance in computing this average.

*Advantages:*

1. Its meaning is clear to the ordinary reader.
2. It is easily computed from numerical data.

3. It lends itself to algebraic treatment. Yule states that this is "by far the most important desideratum." The averages of two or more series can be obtained from the averages of the individual series. For this reason it is useful in the computation of index numbers.

4. It gives weight to all items in direct proportion to their size. This is very useful when the information desired is per capita consumption or production, or wealth regardless of its distribution. There are instances where this is undesirable, as is pointed out later.

5. In some cases the arithmetic average can be found without a knowledge of the individual items. Thus, the per capita annual consumption of sugar, flour, candy, tea, apples, etc., can be determined from a knowledge of the annual production and the total production.

*Disadvantages:*

1. The arithmetic average may not be represented in the actual data. Thus, above, we found the annual average egg production to be 135.98 eggs. Obviously no hen lays a fractional part of an egg. Moreover, there may not have been a hen in the lot that laid either 135 or 136 eggs.

2. Every item has the same weight. This gives in some instances undue weight to the extreme items. Thus a jury assessing damages may have on it two prejudiced jurors, one in favor of the defendant, one an enemy of the defendant. The favorable juryman might desire an assessment of 1¢, the unfavorable juryman might desire $100,000. Ten others might agree that $500 was satisfactory. Evidently an arithmetic average would work an injustice.

3. It cannot be used to advantage with qualitative data, data incapable of numerical measurement. The question here arises as to whether there are any essentially qualitative data. Does not the seeming qualitative character arise from our present inability to measure accurately?

**57. Three theorems.**—Theorem I: *The algebraic sum of the deviations of a series of magnitudes from their arithmetic mean is zero.*

*Proof:*

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{N}$$

$$\therefore N\overline{X} = x_1 + x_2 + \cdots + x_n$$

Hence

$$(x_1 - \overline{X}) + (x_2 - \overline{X}) + \cdots + (x_n - \overline{X}) = 0 \qquad \text{Q.E.D.}$$

Theorem II: *The arithmetic average of two (or more) series can be obtained from the averages of the individual series by means of the formula:*

$$\overline{X} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

*Proof:*

Let

$\overline{X}_1$ = arithmetic average of first series.
$\overline{X}_2$ = arithmetic average of second series.
$\overline{X}$ = arithmetic average of combined series.
$N_1$ = sum of frequencies in first series.
$N_2$ = sum of frequencies in second series.
$N = N_1 + N_2$ = sum of frequencies in combined series.

Then, from the definition of an arithmetic average, we have

$$\overline{X} = \frac{\Sigma x}{N};\ \overline{X}_1 = \frac{\Sigma x_1}{N_1};\ \overline{X}_2 = \frac{\Sigma x_2}{N_2}$$

But

$$\Sigma x = \Sigma x_1 + \Sigma x_2$$

Therefore

$$N\overline{X} = N_1\overline{X}_1 + N_2\overline{X}_2$$

and

$$\overline{X} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2} \qquad \text{Q.E.D.}$$

Theorem III: *The sum of the squares of the deviations from the arithmetic average is a minimum.*

*Proof:*

$$\begin{aligned} Y &= (X - x_1)^2 + (X - x_2)^2 + \cdots + (X - x_n)^2 \\ &= NX^2 - 2X(x_1 + x_2 + \cdots + x_n) + (x_1{}^2 + x_2{}^2 + \cdots + x_n{}^2) \end{aligned}$$

where $X$ represents that value of $x$ from which the deviations have been taken.

Now

$$U = ax^2 + bx + c$$

$$= a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2 - 4ac}{4a}$$

has a minimum value when $x = -b/2a$.
Then $Y$ has a minimum value when

$$X = \frac{x_1 + x_2 + \cdot \cdot \cdot + x_n}{N} = \overline{X}$$

Hence $Y$ has a minimum value when the deviations are taken from $X = \overline{X}$, the arithmetic average.[1]

**Exercises.**

1. Compute the arithmetic average price paid for breakfast, dinner, supper. Compute by the general method, short-cut method, step deviation. Use table I in the appendix.

2. Compute the arithmetic average price paid for all three meals combined. Compute by the general method, short-cut method, step deviation. Compute by combining the averages found in ex. 1.

3. Verify in examples 1 and 2 that the sum of the deviations from the arithmetic average is zero.

4. Compute the arithmetic average egg production for the year (a) 1902–03; (b) 1903–04; (c) 1905–06. Compute by all three methods. Use table II in the appendix.

5. Find the arithmetic average grade of 127 Colorado College freshmen in English. Compute by all three methods. Use table III in the appendix.

6. Compute the arithmetic average earnings per hour for male employees in chemical and lumber industries as given in table V in the appendix.

Show how theorem II can be used to find the arithmetic average of these series.

In computing these averages, for the interval *under 20¢* use 10¢ as the mid-point.

7. Compute the arithmetic average:
   (a) wage paid to bituminous coal miners. Table 2.
   (b) size of farm. Table 9.
   (c) earnings per hr. for laborers inside mines. Table 14.
   (d) income among single women of continental U. S. Ex. 2. §45.

[1] A second proof can be given involving the calculus. A function has a minimum for that value of the variable which causes the first derivative to vanish and gives a positive value to the second derivative. We have

$$Y' = 2(X - x_1) + 2(X - x_2) + \cdot \cdot \cdot + 2(X - x_n) = 0,$$

if

$$X = \frac{x_1 + x_2 + \cdot \cdot \cdot + x_n}{N} = \overline{X}$$

We have further

$$Y'' = 2N > 0$$

**58. Effect of grouping upon the average.**—In computing the arithmetic average, each frequency is multiplied by the measure of the mid-point of the corresponding class interval. This is equivalent to assuming that the arithmetic average of the items within the interval is the mid-point of the interval. The arithmetic average of the items within an interval often does not coincide with the measure of the mid-point, and hence the computed arithmetic average is not the true average. The computed average depends upon the grouping. Different class intervals yield, in general, different arithmetic averages. This is illustrated by the groupings of the data listed in table 39.

**Table 39.—Total Number of Males 16 Years of Age and Over, Foundry and Metal Workers. Rates in Cents per Hour.**

| *Rates ¢* | *Frequency* | | | *Rates ¢* | *Frequency* | | |
|---|---|---|---|---|---|---|---|
| 7– 7.9 | 2 | 4 | 6 | 31–31.9 | 11 | 43 | 63 |
| 8– 8.9 | 2 | | | 32–32.9 | 32 | | |
| 9– 9.9 | 2 | 2 | | 33–33.9 | 14 | 20 | |
| 10–10.9 | 0 | | | 34–34.9 | 6 | | |
| 11–11.9 | 1 | 2 | 3 | 35–35.9 | 18 | 35 | 42 |
| 12–12.9 | 1 | | | 36–36.9 | 17 | | |
| 13–13.9 | 0 | 1 | | 37–37.9 | 2 | 7 | |
| 14–14.9 | 1 | | | 38–38.9 | 5 | | |
| 15–15.9 | 12 | 203 | 418 | 39–39.9 | 2 | 10 | 12 |
| 16–16.9 | 191 | | | 40–40.9 | 8 | | |
| 17–17.9 | 43 | 215 | | 41–41.9 | 1 | 2 | |
| 18–18.9 | 172 | | | 42–42.9 | 1 | | |
| 19–19.9 | 11 | 146 | 268 | 43–43.9 | 1 | 2 | 6 |
| 20–20.9 | 135 | | | 44–44.9 | 1 | | |
| 21–21.9 | 36 | 122 | | 45–45.9 | 3 | 4 | |
| 22–22.9 | 86 | | | 46–46.9 | 1 | | |
| 23–23.9 | 32 | 72 | 253 | 47–47.9 | 0 | 4 | 6 |
| 24–24.9 | 40 | | | 48–48.9 | 4 | | |
| 25–25.9 | 115 | 181 | | 49–49.9 | 0 | 2 | |
| 26–26.9 | 66 | | | 50–50.9 | 2 | | |
| 27–27.9 | 92 | 137 | 223 | | | | |
| 28–28.9 | 45 | | | | | | |
| 29–29.9 | 23 | 86 | | | | | |
| 30–30.9 | 63 | | | | | | |

Class interval 1¢, $\overline{X} = 23.48$
Class interval 2¢, $\overline{X} = 23.30$
Class interval 4¢, $\overline{X} = 23.31$

*Special Report on Employees and Wages, U. S. Census, 1900.*

**59. Weighted arithmetic average.**—Some measurements of a variate may be made under more favorable circumstances than others. It would seem proper to attach more importance to the better measurement in computing an average. This is accomplished by treating the better measurement as equivalent to more than one occurrence of this observed value in a set of equally good measurements.

Thus, if one makes two measurements, $x_1$ and $x_2$, of a given variate and decides that $x_1$ is the better measurement, then one may decide to treat this measurement as equivalent to 3 measurements in a set of 4 equally good measurements. In this case the average is

$$\bar{x} = \frac{3x_1 + x_2}{4}$$

In computing this average we are said to give weights of 3 and 1 to $x_1$ and $x_2$. The computed average is said to be a weighted arithmetic average.

Suppose one makes 4 distinct measurements of a variate, the measurements being $x_1 = 5$, $x_2 = 6$, $x_3 = 5$, $x_4 = 5$. The arithmetic average is

$$\bar{x} = \frac{5 + 6 + 5 + 5}{4} = \frac{3 \times 5 + 6}{4} = \frac{3x_1 + x_2}{4} = 5.25$$

This average is sometimes erroneously referred to as a weighted arithmetic average. This error would appear to arise from the form of the shortened method of computation of the ordinary arithmetic average.

Given an examination of any kind, if the questions are not of equal difficulty, the same number of points are not allowed for the different answers when correct. The number of points allowed for the different correct answers is an attempt at weighting. The weights to be assigned are a matter *to be determined by judgment and experience.*

Many of the examples given to illustrate weighted arithmetic averages appear to the present writers to belong to the ordinary arithmetic average. We give a few such.

In computing an index of the cost of living, it is customary to use as weights for the various prices the quantity consumed. It would seem clear that the quantities consumed represent the frequency with which the given price was paid out by the consumer.

In computing an average retail price for the different cuts of meat, it is customary to use as weights the per cent which each cut bears to the total weight of an average carcass. It would seem clear that these percentages represent the relative frequency with which the stated prices for the different cuts are paid by the general public.

In computing an average price for wheat in specified markets, it is customary to weight the prices with the number of bushels sold in the different markets. The weights are simply the frequencies with which the given prices per bushel are paid.

If one loans at simple interest for one year

$1,000 at 8%
2,000 at 7%
3,000 at 6%

what is the average rate of interest received? The arithmetic average of the three rates, 0.06, 0.07, 0.08 is an incorrect answer. It is customary here to call the correct average a weighted average. The average rate is

$$\frac{1{,}000 \times 8 + 2{,}000 \times 7 + 3{,}000 \times 6}{1{,}000 + 2{,}000 + 3{,}000} = 6\tfrac{2}{3} \text{ per cent}$$

For weights, any numbers can be used which are proportional to the amounts placed at interest. It would seem clear that the amounts at interest at the various rates represent the frequencies with which the different rates are paid on a loan of one dollar, and hence this average should be considered as an ordinary arithmetic average computed for the given frequency distribution.

### Exercises.

1. Compute the average retail price of the following cuts of meat:

| *Cuts* | *Cents per Lb.* | *Per Cent Total Weight* | *Cuts* | *Cents per Lb.* | *Per Cent Total Weight* |
|---|---|---|---|---|---|
| Porterhouse and club | 45 | 8.6 | Miscellaneous....... | 20 | 14 |
| Sirloin........... | 40 | 7.6 | Plate............. | 20 | 13 |
| Rib.............. | 32 | 8.4 | Bones............ | 2 | 10.2 |
| Round and rump.... | 30 | 15.7 | Shrinkage.......... | 0 | 1.3 |
| Chuck........... | 24 | 21.2 | | | |

2. One loans $3,000 at 5 per cent, $2,000 at 7.5 per cent, and $1,000 at 15 per cent. What is the average interest rate?

3. Compute the average price per bu. received by the farmers of the U. S. in 1913 for the following: corn, wheat, oats, barley, rye, potatoes. For data consult table IX, appendix.

4. Compute the average farm price per bu. received for wheat in 1923 in Nebr., Kans., Iowa, Okla., Wis., Minn. For data consult table VIII, appendix.

5. Compute the average farm price per bu. Dec. 1, 1923, for wheat, corn, oats, barley, rye, flaxseed. Use table VII, appendix.

6. Verify the following table. Compute the average rate of interest received. One repays a loan of $100 by 25 equal monthly payments of

| MONTH | *Amount due at beginning of mo.* | RATE, % *per year* | MONTH | *Amount due at beginning of mo.* | RATE, %, *per year* |
|---|---|---|---|---|---|
| 0 | $100 | 6.48 | 13 | 48 | 13.50 |
| 1 | 96 | 6.75 | 14 | 44 | 14.72 |
| 2 | 92 | 7.04 | 15 | 40 | 16.20 |
| 3 | 88 | 7.36 | 16 | 36 | 18.00 |
| 4 | 84 | 7.71 | 17 | 32 | 20.25 |
| 5 | 80 | 8.10 | 18 | 28 | 23.14 |
| 6 | 76 | 8.53 | 19 | 24 | 27.00 |
| 7 | 72 | 9.00 | 20 | 20 | 32.40 |
| 8 | 68 | 9.53 | 21 | 16 | 40.50 |
| 9 | 64 | 10.12 | 22 | 12 | 54.00 |
| 10 | 60 | 10.80 | 23 | 8 | 81.00 |
| 11 | 56 | 11.57 | 24 | 4 | 162.00 |
| 12 | 52 | 12.46 | 25 | 0 | |

Answer: Average rate is 12.4%.

$4.54. Of this amount $4.00 is counted as payment on the principal and 54 cents is counted as interest.

7. Compute the average price per family per lb. paid in 1903 in the United States for the following: rice, sugar, coffee, tea, butter, flour and meal, fresh beef. Use table X, appendix.

8. Same as ex. 6: (a) for the North Atlantic States; (b) for the North Central States. Use table XI, appendix.

9. Compute the arithmetic average egg producion for the years given in table II, appendix. Use thirty eggs as a class interval and compare the results with those obtained by using fifteen eggs as a class interval.

10. Compute the arithmetic average weekly wage for the years given in table VI, appendix. Use one dollar as a class interval and compare the results with those obtained by using fifty cents as a class interval. Omit the first zero frequency for 1900 and the last zero frequency for 1890.

11. Compute the arithmetic average output per man for the data in table XIX, appendix. Use one ton as a class interval and compare with the result when one-half ton is used as a class interval.

**60. Mode.**—The mode is the size of that item which occurs most frequently. The modal size of shoe worn is the size most common. Most apples are now sold in bushel boxes. This is the modal size of package for apples. The most frequent size of can for tomatoes or peaches is a quart can. In a factory the wage which is received by the greatest number is the modal wage. A manufacturer of suits, shoes, stockings, underwear, hats is interested in modal sizes. A transportation company in its advertising would feature its modal performance. The mode is a test of fitness to fill a given position, the performance to be expected.

In determining the average length of life of individuals, the modal length of life is that age at which the greatest number die.

In a frequency table, the modal class is the class interval which has the greatest frequency.

The *apparent* mode is the modal class in a *sample* from the universe under discussion. From the sample there is computed a size of item which is a better representation of the modal size of item in the universe under discussion than is the modal class. This is the *computed* mode. The *true* mode may be still different.

*Computation of the mode.*—The mode is sometimes referred to as an *inspectional* average.

(a) If we have the measure of each item in a sample, then the mode is the size of item which occurs most frequently.

(b) If the measures of the individual items have been replaced by a frequency table with an arbitrary class interval, then the apparent mode is the class interval which has the greatest frequency.

(c) The computed mode is obtained by interpolation in a frequency table. The interpolation is made under the assumption that the position of the mode within the modal class is determined by the relative sizes of the adjacent classes. Under this assumption, the mode is computed from the formula

$$Mo = l + \frac{f_1}{f_{-1} + f_1}i$$

where

*Mo* equals the mode;

$l$, the lower limit of the modal class;

$i$, the class interval;

$f_{-1}$, the frequency of the adjacent class interval below the modal class;

$f_1$, the frequency of the adjacent class interval above the modal class.

The above formula gives a mode computed upward from the lower limit of the modal class. The same numerical result can be obtained from the following formula wherein a mode is obtained, computed downward from the upper limit $L$ of the modal class:

$$Mo = L - \frac{f_{-1}}{f_{-1} + f_1}i.$$

These formulas will be illustrated by means of the data for roving-frame tenders in cotton mills for 1900, Table VI, appendix. The following abstract from that table shows that the modal class is 6.00–6.49.

| *Rates per Week (Dollars)* | *Frequencies* |
|---|---|
| 5.50–5.99 | 36 |
| 6.00–6.49 | 66 |
| 6.50–6.99 | 61 |

$$Mo = 6.00 + \frac{61}{36 + 61} \times 0.50 = 6.31$$

$$Mo = 6.50 - \frac{36}{36 + 61} \times 0.50 = 6.31$$

(d) If the mode is not clearly present in a given class interval, or if the series appears to have more than one mode, the location of the mode may sometimes be determined by a regrouping of the data, using a wider class interval. If, in the regrouping, an item of a given size is always found in the group of maximum frequency, then the size of this item is taken to be the modal size.

| *Size of Item (Mid-point)* | *f* | | | | | |
|---|---|---|---|---|---|---|
| 1 | 18 | 40 | | 88 | 65 | |
| 2 | 22 | | 47 | | | 90 |
| 3 | 25 | 48 | | | | |
| 4 | 23 | | 43 | | 64 | |
| 5 | 20 | 41 | | 82 | | |
| 6 | 21 | | 46 | | | 76 |
| 7 | 25 | 41 | | | 56 | |
| 8 | 16 | | 30 | | | |
| 9 | 14 | | | | | |

Thus, in the above hypothetical example, the item whose size is 3 is taken to be the modal size of item. It should be remarked here that, if more than one mode seems to appear, this may be due to one or the other of two causes. First, the bimodal character may arise from the small size of the class interval. In this case, a true mode will usually appear upon widening the interval properly. The apparent bimodal character will also disappear upon taking a sample containing a greater number of items. If the bimodal character does not disappear upon the application of one or the other of these methods, then more than one mode truly exists. In this case, one should examine the universe from which the sample is taken in order to determine whether it is homogeneous. The bimodal character may arise from the fact that the universe under discussion consists of two groups, each with a distinct mode different for the different groups.

(e) Construct a smoothed frequency graph from the given data. The abscissa of the highest point on the curve is taken to be the modal size.

(f) Construct a cumulative frequency graph. The abscissa of the point where the curve is steepest is taken to be the modal size. For at that point a given increase in the size of the items produces a greater change in the frequency than at any other point of the curve. Thus, for the data in table 28, plotted in fig. 22, the modal size of farm is approximately 40 acres.

**61. Advantages and disadvantages.**—This average is easily comprehended. Items of extreme measure, either large or small, have no effect upon the mode provided they are not in the modal class. The size and number of the extreme items need not be known provided it is known that they are not in the modal class. The mode is easily found.

The mode is not well adapted to numerical computations in the combinations of series. A clearly defined mode does not always exist. The fact that the mode is not influenced by the relatively small or large items is a disadvantage in those cases where it becomes necessary so consider all items.

### Exercises.

Determine the apparent mode and computed mode in the following cases (References are to tables in the appendix):

1. From American Experience Mortality Table determine the average length of life. Table XII.

2. Expenditures for meals. Table I.
3. Roving-frame tenders in cotton mills, 1890. Table VI.
4. Annual egg production. Table II.
5. Distribution of grades of 127 Colorado College freshmen. Table III.
6. Incomes among single women of Continental U. S. Table IV.
7. Earnings of male employees in the U. S. in chemical and lumber industries. Table V.
8. Relative price of 1,437 commodities in 1918. Table XIII.

**62. Median, definitions.**—If the items of a series are arranged in the order of their magnitudes, the measure of the central item in the series is termed the *median*. When all of the items are given in an ungrouped form but arranged in the order of their magnitudes, the central item is determined by the simple process of counting, beginning at either end, up to that item such that there are an equal number of items on either side of it.

For a series containing an odd number of items, the median is the measure of the $\frac{N+1}{2}$th individual of the group. If a series contains an even number of items, it is customary to express the median as the arithmetic average of the measures of the two central items, even though there may be no item in the series with this measure.

Thus, the median of the numbers, 2, 4, 6, 8, 10 is 6, a number of the series; while the median of the numbers 2, 3, 6, 8, 10, 12 is 7, the arithmetic average of the two central numbers 6 and 8, even though 7 is not a member of the series.

In speaking of the sales of a salesman, the mode emphasizes the typical performance, ignoring exceptionally good days and very poor days. The median takes into account all sales on all days, giving all the same importance.

This would be a fair average for a jury to use in assessing damages.

This would be the average used in mass production in engineering, for example when speaking of the sizes of bolts.

A median age of life would be the age at which exactly half of a given group were still living.

**63. Computation.**—Let us consider the hypothetical data below:

| *Class Interval* | *Frequency* |
|---|---|
| 0 to 10 | 1 |
| 10 to 20 | 3 |
| 20 to 30 | 5 |
| 30 to 40 | 2 |
| *Total* | 11 |

The number of the median item is 6. This item occurs in the group of class interval 20 to 30. In the absence of the original data, we do not know how these 5 items are distributed within this interval. For the purposes of the computation it is assumed that the measures in this class interval are uniformly distributed throughout the interval.

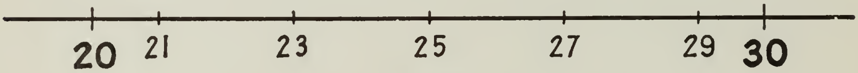


Fig. 32.

This assumption is not necessary for any interval except the one within which the median lies. Since there are 5 items within a class interval whose width is 10, by the above assumption these items are separated by an interval of 2 units. If we assume further that the first and last items within the interval are one-half of this uniform distance, that is, a distance of one unit, from the end points, the distribution within the interval is as shown graphically in fig. 32. The median is the measure of the sixth item. But there are 4 items before we reach the measure 20. Then the median is the measure of the second item within this class interval. Thus the median is 23.

In general terms, the formula for the measure of the median item is

$$Md = l + \frac{\frac{N}{2} - F}{f} i$$

$$= L - \frac{\frac{N}{2} - F'}{f} i$$

where

$Md$ equals the median;

$l$, the lower limit of the class interval which contains the median;

$L$, the upper limit of the class interval which contains the median;

$f$, the frequency in the class interval which contains the median;

$i$, the class interval;

$N$, the total frequency;

$F$, sum of frequencies in all classes below $l$;

$F'$, sum of frequencies in all classes above $L$.

Applying the formula to the above hypothetical data, we have

$$Md = 20 + \frac{\frac{11}{2} - 4}{5} \times 10 = 23$$

**64. Graphical methods.**—(a) *Interpolation on a cumulative frequency graph.* From the mid-point of an ordinate which represents the total frequency, draw a horizontal line. From the point where this horizontal line cuts the cumulative graph, drop a perpendicular to the horizontal axis. The reading at the foot of this perpendicular is the value of the median as obtained from this graph. These remarks are illustrated in fig. 22, where the average (median) size of farm is determined as approximately 80 acres.

Draw two cumulative graphs, one on the more than basis, the other on the less than[2] basis. From the point of intersection of these two graphs drop a perpendicular to the horizontal axis. The reading at the foot of this perpendicular is the median. This is illustrated in fig. 22.

(b) *Frequency graph.* The reading at the foot of the ordinate which bisects the *area* under the frequency graph is the *median,* for the total area under the frequency graph represents the total frequency.

**65.** Theorem IV: *The sum of the deviations from the median, all considered positive, is a minimum.*

[2] For a description of these terms see article 47.

*Proof:* The sum of the distances from two given points is the same for any point between them and less than for any external point.

For any array, the median has as many points on one side as on the other. Hence these points can be paired and the median is between each pair, and our theorem is proved.

*Illustration:* The set of numbers 2, 5, 7, 12, 18 has 7 for a median. The sum of the deviations from the median, all considered positive, is $5 + 2 + 0 + 5 + 11 = 23$. The sum of the deviations from 8, all considered positive, is $6 + 3 + 1 + 4 + 10 = 24$.

**66. An approximate relation.**—If the frequency graph is symmetrical, then $\overline{X} = Mo = Md$.

If, however, the frequency graph is not symmetrical and the departure from symmetry is not marked, then it has been observed that the following relation is approximately satisfied:

$$Mo = \overline{X} - 3(\overline{X} - Md)$$

Another form of this equation is

$$Md = Mo + \tfrac{2}{3}(\overline{X} - Mo)$$

That is, the median is $\frac{2}{3}$ of the distance from the mode toward the arithmetic average. This is illustrated in fig. 33.

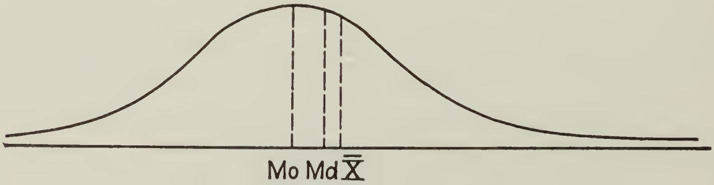


Fig. 33.

**67. Time series.**—Let us consider the following data which give, by months, the bituminous coal produced in the United States in 1922 as given by the United States Geological Survey, weekly report no. 298.

**Table 40.—Bituminous Coal Produced by Months in the United States in 1922.**

| *Month* | *Production, 1,000 Tons* | *Month* | *Production, 1,000 Tons* |
|---|---|---|---|
| Jan | 37,489 | July | 17,147 |
| Feb | 40,856 | Aug | 27,538 |
| March | 49,976 | Sept | 39,413 |
| Apr | 16,000 | Oct | 44,907 |
| May | 20,601 | Nov | 45,103 |
| June | 22,624 | Dec | 46,240 |
| | | Total | 407,894 |

July 1 is the median with respect to time. This median with respect to time is important. At noon one takes account of what has already been accomplished and compares it with what one desires to accomplish for the entire day. Wednesday evening one compares the three days' work already done with what one desires to accomplish for the week. This median date is useful in planning for the future. With respect to the production of coal, the number of tons produced up to July 1 would give some indication to the operator, when compared with the production for previous years, as to what the production would be for the current year.

The total production for the year was 407,894,000 tons. Half of this is 203,947,000 tons. If one cumulates the production for the year, one finds that by the end of June the production has reached 187,546,000 tons. Computing the median date from the formula, one finds

$$\begin{aligned}\text{Date of half of total production} &= \text{June } 30 + \frac{16{,}401}{17{,}147} \times 31 \\ &= \text{July } 29.6\end{aligned}$$

This would be about noon on July 30. With coal, this date would be better than July 1 as a date to use in forecasting the total yearly production.

In general, this is the time meant when one uses the term "median time" with respect to historical data. If any other median is used, one should specify in detail just what is meant.

A third median which is sometimes used gives, with respect to the data on coal above, that month in which the production was an average. That is, one arranges the productions per month in the order of their magnitudes and determines the median of this array. In this instance, since there is an even number of items, one takes one-half of the sum of the productions in the sixth and seventh items in the array (in this case the productions for January, 37,487,000 tons, and September, 39,413,000 tons), giving an average of 38,477,000 tons as the median monthly production. Using this average, a producer would say that the production for the month was up to the average, or below or above the average, as the case might be.

**68. Advantages and disadvantages.**

Among the advantages of the median note that:

1. It is easy to determine.

2. A median can be determined when the class interval for the extremes is open at one end, provided the frequency is given. Thus, in giving the frequency data for incomes the upper class interval may read $1,000,000 and over. The lower interval may read $100 and under.

3. The median can be determined for data in which the class intervals are unequal.

4. It is adapted to series in which the central items are bunched although there does not exist a well defined mode.

5. It is not distorted by a few items exceptionally large or small.

6. The size of the median can never be changed greatly by the addition of a few more items.

Thus the series

$$1, 2, 3, 7, 7, 12, 20 \text{ has } Md = 7; Mo = 7$$

while the series

$$1, 2, 3, 7, 7, 12, 20, 20, 20 \text{ has } Md = 7; Mo = 20.$$

Among its disadvantages we note that

1. It does not lend itself to algebraic treatment. That is, one cannot compute the median of several series by combining the medians of the component series. Thus

1, 2, 3, 4, 5 has 3 for a median;
11, 18, 19, 20, 41 has 19 for a median.

The median of 3 and 19 is 11, while the median of 1, 2, 3, 4, 5, 11, 18, 19, 20, 41 is 8.

2. The above example illustrates that the median may not be a measure of any item in the sample. The median may not be the measure of any item in the universe from which the sample is taken.

3. The median may be located at a point in the series where there are few or no items and thus not represent the typical state of affairs. Thus in the series

1, 2, 5, 5, 5, 5, 7, 11, 13, 13, 13, 13, 15, 18

the median is 9, which is not a member of the series. The series

1, 2, 5, 5, 5, 5, 7, 9, 11, 13, 13, 13, 13, 15, 18

also has 9 as a median but it does not appear to be typical. This state of affairs is likely to be true for a bimodal series.

### Exercises.

Compute the median in the following cases (References are to tables in the appendix):

1. Median length of life from the American Experience Mortality Table. Table XII.
2. Expenditure for food. Table I.
3. Annual egg production. Table II.
4. Income among single women of the Continental United States. Table IV.
5. Earnings of male employees in chemical and lumber industries. Table V.
6. Roving-frame tenders in cotton mills. Table VI.
7. Relative prices of 1,437 commodities. Table XIII.

**69. Geometric average, definition and computation.**—The *geometric average* of the measures of $n$ items is the $n$-th root of the product of these $n$ measures. Thus, if we let $G$ represent the geometric average, we have

$$G = \sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdot \cdot \cdot x_{N-1} \cdot x_N}$$

As an illustration, the geometric average of 4, 8, 16 is 8 for $\sqrt[3]{4 \cdot 8 \cdot 16} = 8$.

It is obvious that if any measure is 0, then $G = 0$.

The actual computation of the geometric average is facilitated by the use of logarithms. For, if we take the logarithm of both sides of the equation, we have

$$\log G = \frac{\log x_1 + \log x_2 + \cdots + \log x_N}{N}$$

That is, we find $\log G$ as the arithmetic average of the logarithms of the given measures, whence $G$ can be found.

If the measurements are given in a frequency table, where $x_1$ has a frequency $f_1$, $x_2$ a frequency $f_2$, etc., then

$$G = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdots x_r^{f_r}}$$

where

$$N = f_1 + f_2 + \cdots + f_r$$

When the measures are to be weighted, the assumed weights take the place of the frequencies in a frequency table in which each individual measure is considered of the same importance. Thus, for a weighted geometric average we have

$$G = \sqrt[N]{x_1^{w_1} \cdot x_2^{w_2} \cdots x_r^{w_r}}$$

where

$$N = w_1 + w_2 + \cdots + w_r$$

and $w_1, w_2, \ldots, w_r$ are the weights assigned to the measures $x_1, x_2, \ldots, x_r$, respectively. When logarithms are used, this formula becomes

$$\log G = \frac{w_1 \log x_1 + w_2 \log x_2 + \cdots + w_r \log x_r}{N}$$

The method of computation by the use of logarithms may be illustrated by table 41, which is a selected list from table XIII, appendix, which gives the distribution of the relative prices of 1,437 commodities in 1918.

**Table 41.**—A Selected List from Table XIII in Appendix, on the Relative Price of 1,437 Commodities.

| *Relative Price* | *Mid-point: x* | *f* | *Log x* | *f · log x* |
|---|---|---|---|---|
| 50– 69 | 60 | 4 | 1.77815 | 7.11260 |
| 70– 89 | 80 | 17 | 1.90309 | 32.35253 |
| 90– 109 | 100 | 61 | 2.00000 | 122.00000 |
| 110– 129 | 120 | 64 | 2.07918 | 133.06752 |
| 130– 149 | 140 | 130 | 2.14613 | 278.99690 |
| 150– 169 | 160 | 212 | 2.20412 | 467.27344 |
| 170– 189 | 180 | 219 | 2.25527 | 493.90413 |
| 587 | 587 | 1 | 2.76864 | 2.76864 |
| 900 | 900 | 1 | 2.95424 | 2.95424 |
| 2,049 | 2,049 | 1 | 3.31154 | 3.31154 |
| 3,009 | 3,009 | 1 | 3.47842 | 3.47842 |
|  | Totals | 711 |  | 1,547.21996 |

Substituting in the formula, we find

$$\log G = \frac{1547.21996}{711} = 2.17612$$

Whence

$$G = 150$$

**70. Properties of geometric average.**—The geometric average is used in averaging rates or ratios rather than quantities. It is used in averaging rates of increase. For example, it is used in finding the average rate of increase in skill, in writing, in language work, in marksmanship, in golf, etc. It is used in computing the average rate of increase in population of a nation or city.

A geometric average is used in computing an index number of prices, for in computing an index number, rates of change of prices are of major importance. A rise in price of one article from 60 to 90 represents the same rate of change of price as a rise in price of another article from 200 to 300. Namely, in both cases we have a 50 per cent increase in price. If two articles of the same relative importance have price changes, the one from 100 to 25, the other from 100 to 400, then the geometric average price of the two articles has not changed, for

$$\sqrt{25 \times 400} = 100$$

But the arithmetic average price shows a decided increase in price, for $\frac{1}{2}(25 + 400) = 212.5$. This price of 212.5 is incorrect as a measure of the average rate of change in price.

The geometric average is used whenever one has a limited lower range and an unlimited upper range. One finds this condition to exist for data with respect to incomes. Here the lower limit is zero and the upper range is unlimited. Indeed it becomes impractical to plot data with respect to incomes on a natural arithmetic chart. A logarithmic chart is desirable.

The geometric average is generally used in computing average prices of stocks. A stock cannot decline more than 100 per cent below par value, but may increase in value 1,000 per cent or more above par value. We have here an unlimited upper range, an excessive range.

In general, whenever the upper range is excessive and the frequency curve when plotted is not symmetrical, plot the logarithms of the measures as abscissas against the frequencies as ordinates. If the resulting curve tends to symmetry, this indicates that a geometric average of the original data is the best average to use to typify the aggregate. For we have seen that for a symmetrical distribution an arithmetic average is indicated, and an arithmetic average of the logarithms of the measures is equivalent to a geometric average of the original measures.

If in a time series the measures increase with increasing time and when plotted give a curve of the exponential or compound interest type, a geometric average rate of increase of the measures is indicated.

Thus, if $p_o$ represents the population at the beginning of a period and $r$ is the average rate of increase of the population, then at the end of $n$ years the population $p_n$ would be given by the equation

$$p_n = p_o(1 + r)^n$$

Hence, if we know the population at the beginning and end of a period of $n$ years, we can compute from this equation the average rate $r$ of increase (or decrease), for

$$r = \sqrt[n]{\frac{p_n}{p_o}} - 1$$

Thus, if the population of the United States increases in 10 years from 75,994,575 to 91,972,266, we have

$$\begin{aligned} r &= \sqrt[10]{\frac{91{,}972{,}266}{75{,}994{,}575}} - 1 \\ &= \sqrt[10]{1.21025} - 1] \\ &= 1.019 - 1 = 0.019 \end{aligned}$$

That is, the rate of increase was approximately 2 per cent per year.

Again, if $1,000 increases at compound interest to $2,000 in 12 years, there has been an increase of 100 per cent.

The arithmetic average is 8⅓ per cent. The geometric average rate at which the money increased is given by the equation

$$\begin{aligned} r &= \sqrt[12]{\frac{2{,}000}{1{,}000}} - 1 \\ &= 1.06 - 1 = 0.06, \text{ or } 6\% \end{aligned}$$

**71. Theorems.**—Theorem I: *If the ratios of the geometric average to the measures which it exceeds or equals be multiplied together, the product will equal the product of the ratios of the geometric average to those measures which exceed it in value.*

*Proof:* For an arithmetic average (disregarding sign) the sum of the deviations of the measures below the average equals the sum of the deviations of the measures above the average. Apply this truth to the equation

$$\log G = \frac{\log x_1 + \log x_2 + \cdots + \log x_N}{N}$$

Assume that $\log x_1, \log x_2, \ldots, \log x_K$ are less than or equal to $\log G$ and that $\log x_{K+1}, \ldots, \log x_N$ are greater than $\log G$. We have

$$\begin{aligned} (\log G - \log x_1) + (\log G - \log x_2) + \cdots + (\log G - \log x_K) \\ = (\log x_{K+1} - \log G) + \cdots + (\log x_N - \log G) \end{aligned}$$

Whence

$$\log\left(\frac{G}{x_1}\cdot\frac{G}{x_2}\cdot\cdot\cdot\cdot\cdot\frac{G}{x_K}\right) = \log\left(\frac{x_{K+1}}{G}\cdot\cdot\cdot\cdot\cdot\frac{x_N}{G}\right)$$

or

$$\frac{G}{x_1}\cdot\frac{G}{x_2}\cdots\frac{G}{x_K} = \frac{x_{K+1}}{G}\cdots\frac{x_N}{G} \qquad \text{Q.E.D.}$$

*Illustration:*

$$\sqrt[5]{2 \cdot 4 \cdot 6 \cdot 9 \cdot 18} = 6$$

$$\tfrac{6}{2} \cdot \tfrac{6}{4} \cdot \tfrac{6}{6} = \tfrac{9}{6} \cdot \tfrac{18}{6}$$

Theorem II: *The arithmetic average of n positive quantities is greater than their geometric average.*

*Proof:* Consider the geometric average $(x_1\ x_2\ .\ .\ .\ x_N)^{\frac{1}{N}}$. If $x_1, x_2, \ldots$ be not all equal, replace the greatest and least of them, say $x_1$ and $x_N$, by $\frac{1}{2}(x_1 + x_N)$.

Now

$$(x_1 - x_N)^2 > 0$$

Add $4x_1x_N$ to both sides of this inequality.

We have

$$x_1^2 + 2x_1x_N + x_N^2 > 4x_1x_N$$

Whence

$$\tfrac{1}{4}(x_1^2 + 2x_1x_N + x_N^2) > x_1x_N$$

and

$$\tfrac{1}{2}(x_1 + x_N) > \sqrt{x_1x_N}$$

But $\sqrt{x_1x_N}$ is the geometric average of the original items while $\frac{1}{2}(x_1 + x_N)$ is the geometric average of the two equal items $\frac{1}{2}(x_1 + x_N)$ which replace $x_1$ and $x_N$. Hence the result has been to increase the geometric average while the arithmetic average has remained unaltered. If the new set of $N$ quantities be not all equal, replace the greatest and least as before, and so on. By repeating this process sufficiently often, we can make all the quantities as nearly equal as we please, and then the geometric average becomes equal to the arithmetic average. But, since the latter has remained unaltered throughout, and the former has increased at each step, it follows that the first geometric average, namely

$$\sqrt[N]{x_1x_2\ \cdot\ \cdot\ \cdot\ x_N}$$

is less than the first arithmetic average, namely

$$\frac{1}{N}(x_1 + x_2 + \cdot\cdot\cdot + x_N) \qquad \text{Q.E.D.}$$

*Illustration:*

$$(1 \cdot 3 \cdot 5 \cdot 9)^{\frac{1}{4}} < (5 \cdot 3 \cdot 5 \cdot 5)^{\frac{1}{4}} < (5 \cdot 4 \cdot 4 \cdot 5)^{\frac{1}{4}}$$
$$< (4 \cdot 5 \times 4 \cdot 5 \times 4 \cdot 5 \times 4 \cdot 5)^{\frac{1}{4}} = 4 \cdot 5$$

while

$$\tfrac{1}{4}(1 + 3 + 5 + 9) = 4 \cdot 5$$

Therefore

$$(1 \cdot 3 \cdot 5 \cdot 9)^{\frac{1}{4}} < \tfrac{1}{4}(1 + 3 + 5 + 9)$$

Theorem III: *The geometric average of the ratios of corresponding observations in two series is equal to the ratio of their geometric averages.*

*Proof:* For, if

$$x_o', x_o'', x_o''', \cdot \cdot \cdot , x_o^N$$

represent the observations of $N$ items in the first series and

$$x_1', x_1'', x_1''', \cdot \cdot \cdot , x_1^N$$

represent the corresponding observations of $N$ items in the second series, then

$$G = \left(\frac{x_1'}{x_o'}\,\frac{x_1''}{x_o''}\,\frac{x_1'''}{x_o'''} \cdot \cdot \cdot \frac{x_1^N}{x_o^N}\right)^{\frac{1}{N}}$$

$$= \frac{(x_1'x_1''x_1''' \cdot \cdot \cdot x_1^N)^{\frac{1}{N}}}{(x_o'x_o''x_o''' \cdot \cdot \cdot x_o^N)^{\frac{1}{N}}} = \frac{G_1}{G_o} \qquad \text{Q.E.D.}$$

$G_1$ and $G_o$ represent the geometric averages of the separate series.

In particular, the first set of observations may represent the prices of $N$ commodities in the base or zero year, while the second set represents the prices of the same $N$ commodities in some other year. In this case, $G$ is an index number of prices for the second year, using the zero year as base.

Theorem IV: *The geometric average of the series formed by combining r different series each with the same frequency is the geometric average of the geometric averages of the separate series.*

*Proof:* Let there be $r$ series each with $N_1$ items.

Let $G_1, G_2, \ldots , G_r$ represent the geometric averages of the separate series, while $G$ represents the geometric average of the series formed by combining the $r$ different series.

Let

$$x_{1i},\ x_{2i},\ x_{3i},\ \cdot \cdot \cdot ,\ x_{N_1 i}$$

represent the $N_1$ items in the $i$-th series.

We have

$$G_1 = \sqrt[N_1]{x_{11}x_{21} \cdots x_{N_1 1}}$$

$$G_2 = \sqrt[N_2]{x_{12}x_{22} \cdots x_{N_1 2}}$$

$$\cdots\cdots\cdots\cdots$$

$$G_r = \sqrt[N_r]{x_{1r}x_{2r} \cdots x_{N_1 r}}$$

$$G = \sqrt[rN_1]{(x_{11}x_{21} \cdots x_{N_1 1})(x_{12}x_{22} \cdots x_{N_1 2}) \cdots (x_{1r}x_{2r} \cdots x_{N_1 r})}$$

$$= \sqrt[r]{G_1G_2 \cdots G_r} \qquad \text{Q.E.D.}$$

## 72. Advantages and disadvantages.—

*Advantages:*

1. It gives equal weight to equal rates of change. Hence it is an appropriate average to use in averaging rates of price change.

2. It lends itself to algebraic manipulation. This is evident from theorems III and IV.

3. All items are used in its computation and hence have an influence upon the result.

4. The geometric average is not influenced as much by extreme deviations as is the arithmetic average.

*Disadvantages:*

1. It is comparatively difficult to compute.

2. It cannot be used for series in which the end class intervals are left open or indeterminate.

3. A knowledge of the measure of every item is necessary.

### Exercises.

1. Compute the geometric average of 3, 6, 12, 24, 48.

2. The following data give the population of the designated countries at two different times. Find the yearly rate of increase (*Source: "International Encyclopaedia"*).

| | | |
|---|---|---|
| France.............. | 1891— 38,343,192; | 1911— 39,601,509. |
| Germany............ | 1890— 49,428,000; | 1910— 64,925,993. |
| United Kingdom..... | 1891— 37,732,922; | 1911— 45,221,615. |
| United States........ | 1890— 62,947,714; | 1910— 91,972,266. |
| Russia.............. | 1897—129,209,297; | 1911—167,003,400. |
| Sweden............. | 1890— 4,785,000; | 1900— 5,136,000. |
| Spain............... | 1887— 17,565,632; | 1910— 19,995,446. |
| Italy................ | 1862— 25,000,000; | 1911— 34,671,377. |

3. If interest is compounded annually, find the rate
   (a) if \$1,000 amounts to \$4,801 in 40 years. *Ans.* 4 per cent.
   (b) if \$1,000 amounts to \$5,516 in 35 years. *Ans.* 5 per cent.
   (c) if \$1,000 amounts to \$1,999 in 9 years. *Ans.* 8 per cent.

4. Compute the geometric average of the relative prices of the 1,437 commodities in table XIII, appendix. *Ans.* 198.3.

**73. Harmonic average, definition and computation.**—The harmonic average of a set of measurements is the reciprocal of the arithmetic average of the reciprocals of the individual measurements. Thus, if we represent the individual measurements by $x_1$, $x_2$, . . . , $x_N$, the formula for the harmonic average $H$ is

$$H = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}$$

If the measurements are given in a frequency series

$$H = \frac{N}{\Sigma f \frac{1}{x}}$$

For two measures $x_1$ and $x_2$,

$$H = \frac{2x_1x_2}{x_1 + x_2}$$

For three measures, $x_1$, $x_2$, and $x_3$,

$$H = \frac{3x_1x_2x_3}{x_1x_2 + x_2x_3 + x_3x_1}$$

The harmonic average must be used in the averaging of time rates. The following examples will illustrate its use.

The unit of work in bricklaying is the brick. Let us suppose that $A$ lays 1 brick each 1.20 minutes; and $B$ lays 1 brick each 0.96 minutes. What is their average rate of laying bricks, and how long will it take them to lay 40,000 bricks?

An arithmetic average of these rates is 1.08 minutes in which to lay 1 brick. One man working at this rate will use 40,000 × 1.08 = 43,200 minutes in laying 40,000 bricks. Two men will require 21,600 minutes or 45 days of 8 hours each. This result is not correct.

If $A$ lays 1 brick in 1.2 minutes, he will lay 400 bricks in an 8-hour day. If $B$ lays 1 brick in 0.96 minutes, he will lay 500 bricks in an 8 hour day. Together they will lay 900 bricks per day. Working together, they will require 40,000 ÷ 900 = $44\frac{4}{9}$ days to lay 40,000 bricks. This is the correct answer. Working together, they lay an average of 450 bricks apiece each day or 1 brick each $1.06\frac{2}{3}$ minutes. This is the correct average rate for doing the given unit of work. This rate is the harmonic average of the given rates, for

$$1.06\tfrac{2}{3} = \frac{2}{\dfrac{1}{1.2} + \dfrac{1}{0.96}}$$

In general, if we use a stop watch to find the time required to do a unit of work in a factory by different individuals, then we must use the harmonic average in computing the average output per day.

### Exercises.

1. Compute the harmonic average of the measures 2, 3, 4. *Ans.* 2.8.

2. *A* does a unit of work in 6 minutes. *B* does a unit of work in 5 minutes. What is their average rate of working? *Ans.* $5\frac{5}{11}$ min.

3. *A* can pick a quart of strawberries in 6 minutes, *B* in 5 minutes, *C* in 4 minutes. What is their average rate of work and how long will it take them to pick 222 quarts? *Ans.* $4\frac{32}{37}$ minutes; 6 hours.

4. *A* can solve a problem in 4 minutes, *B* in 5 minutes, *C* in 6, *D* in 10, and *E* in 12. How many problems can they solve together in 8 hrs.? *Ans.* 384.
What is the average time required to solve a problem? *Ans.* 6.25 minutes.

5. *A* can do a piece of work in 7 days; *B* in 9 days. How long will it take them, working together, to do the job? *Ans.* $3\frac{15}{16}$ days.

In the *Colorado Springs Gazette* for Nov. 21, 1925, the following quotations were given:

6. On apples in lbs. for 25¢:

| | | | |
|---|---|---|---|
| Fancy Jonathan | 4 lbs. | Winesap | 7 lbs. |
| Rome Beauty | 5 lbs. | Missouri Pippin cooking | 10 lbs. |

Compute the harmonic average number of lbs. for 25¢.

7. On vegetables in lbs. for 25¢:

| | | | |
|---|---|---|---|
| carrots | 10 lbs. | white potatoes | 5 lbs. |
| onions | 8 lbs. | sweet potatoes | 4 lbs. |

Compute the harmonic average number of lbs. for 25¢.

8. In pounds for one dollar:

| | | | |
|---|---|---|---|
| English walnuts | 3 lbs. | potatoes | 20 lbs. |
| lard | 6 lbs. | sugar | 15 lbs. |

Compute the harmonic average number of lbs. for $1.

9. In each of the following compute the harmonic average price per pound:

In pounds for 25¢:

| | | | |
|---|---|---|---|
| lima beans | 1 lb. | buckwheat | 3 lbs. |
| peas | 2 lbs. | hominy | 5 lbs. |

In pounds for 25¢:

| | | | |
|---|---|---|---|
| pigs' feet | 3 lbs. | hens | 1 lb. |
| veal stew | 3 lbs. | pork loin | 1 lb. |

10. Three ships make the same round trip in 20, 24, and 30 days respectively. What is the average number of days required to make the round trip? *Ans.* 24 days.

# Chapter VIII

## DISPERSION AND SKEWNESS

**74. Dispersion.**—Let us suppose that the following numbers represent the number of eggs per year laid by individual hens in two flocks of five each.

**Table 42.—Number of Eggs per Year Laid by Individual Hens in Two Flocks of Five Each.**

| *Flock I* | *Flock II* |
|---|---|
| 140 | 50 |
| 145 | 100 |
| 150 (average) | 150 (average) |
| 155 | 200 |
| 160 | 250 |

The average number of eggs per year per hen is the same for both flocks. To state this average does not sufficiently characterize the flock. It is obvious that there is a greater uniformity of production in the first flock.

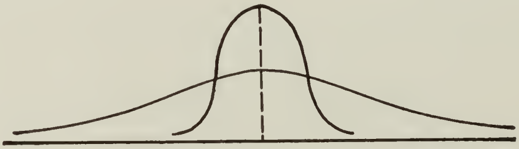


**Fig. 34.**

Two frequency distributions may have the same average, but the distribution of the items about this average may be quite different in the two series. If the sum of the frequencies near the average is a relatively large percentage of the total frequency, the group is said to be *uniform.* Thus one drove of hogs may average 200 lbs. in weight with no hog under 190 or over 210 lbs. in weight. Another drove may also average 200 lbs. but have some pigs weighing 50 lbs. and others weighing 400 lbs. The two curves in fig. 34 show such distributions.

Two frequency distributions may have different averages, while the distribution of the items about the average may be the same in the two groups. This is illustrated by the following hypothetical frequency distributions.

**Table 43.—Hypothetical Frequency Distribution; Weights of Turkeys and Cattle.**

| *Frequency* | *Turkeys* | *Cattle* |
|---|---|---|
| | *Weight, Lbs.* | *Weight, Lbs.* |
| 1 | 6 | 798 |
| 3 | 7 | 799 |
| 5 | 8 (average) | 800 (average) |
| 3 | 9 | 801 |
| 1 | 10 | 802 |

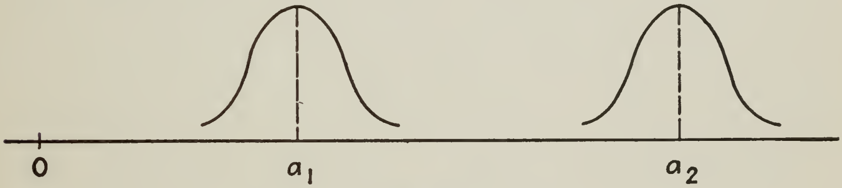


**Fig. 35.**

The curves in fig. 35 show two such distributions.

Two frequency distributions may have different averages and also different distributions of the items about the average.

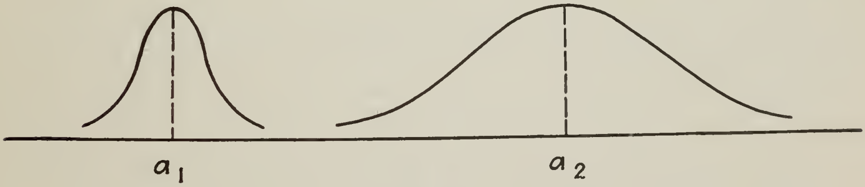


**Fig. 36.**

The curves in fig. 36 show such distributions.

It is clear that there is needed some measure of the degree with which the items are concentrated about the average or deviate from it. Such a measure is called a *measure of dispersion.*

Measures of dispersion which are easy to compute are:

1. The range.
2. The quartile deviation.

A good measure of the variation of the items from a mean of the distribution is obtained by computing an average of these deviations. With this viewpoint one computes usually:

3. The Average Deviation from the:
   (a) Median.
   (b) Mode.
   (c) Arithmetic average.
4. The Standard Deviation.

**75. The range.**—The *range* is the difference between the smallest and largest items of the distribution. It is easily found. For the first flock of chickens in the previous article the range is 20. For the second flock the range is 200. The range is commonly used in quoting interest rates and security prices. Ordinarily the range gives a rough measure of dispersion to which no great significance should be attached. Its value depends upon only two items in the distribution.

**76. The quartile deviation—Deciles—Percentiles.**—The quartiles are the magnitudes which divide the distribution into four equal parts. Below the first, or lower, quartile are one fourth of the items. Below the second quartile, or median, are one-half of the items. Below the third, or upper, quartile are three-fourths of the items.

The quartile deviation ($Q$) is found by taking half the difference between the third quartile ($Q_3$) and the first quartile ($Q_1$):

$$Q = \frac{Q_3 - Q_1}{2}$$

The quartile deviation is easily computed and is a better measure of dispersion than the range. The computation is a matter of interpolation, under the assumption that the magnitudes are evenly distributed within each class. For turkey weights the computation is as follows:

The frequency is 13. One-fourth of this is $3\frac{1}{4}$. We desire the weight of the turkey whose number is 3.25. This turkey is in the class interval 6.5 lbs. to 7.5 lbs. The frequency is 3. The turkey's weight $Q_1$ is then

$$Q_1 = 6.5 + \frac{2.25}{3} \times 1 = 7.25 \text{ lbs.}$$

In like manner

$$Q_3 = 8.5 + \frac{\frac{3}{4}}{3} \times 1 = 8.75 \text{ lbs.}$$

Hence,

$$Q = \frac{8.75 - 7.25}{2} = 0.75 \text{ lbs.}$$

The *deciles* are those values of the variable which divide the entire frequency into ten equal parts. Deciles are not used in computing a numerical measure of dispersion. A statement of the range for each decile aids in forming a picture of the extent and character of the dispersion.

*Percentiles* are those values of the variable which divide the entire frequency into one hundred equal parts. A statement of the range for each percentile, with the common frequency, gives a more detailed picture of the dispersion than is given by the deciles or quartiles. This multiplicity of detail is rarely helpful and is seldom used.

The number of items which falls within the first decile is $\frac{N}{10}$, within the first percentile $\frac{N}{100}$, where $N$ is the total number of items.

Prof. Wesley C. Mitchell[1] makes use of deciles and percentiles to show graphically the dispersion from year to year in price changes. One graphic device used by Prof. Mitchell is to compute the deciles for each year and draw a continuous curve for each. A second graph is made as

[1] W. C. Mitchell, "Index Numbers of Wholesale Prices in the United States and Foreign Countries," U. S. Bureau of Labor Statistics, Bulletin No. 284, pp. 14–15. W. C. Mitchell, "Business Cycles," University of California Studies, Berkeley, 1913, p. 112.

For adaptations from Mitchell see H. Jerome, "Statistical Methods," p. 153, and H. Secrist, "An Introduction to Statistical Methods," 1925, pp. 331–335.

follows: for each year there is plotted a vertical line whose length represents the range of prices for the commodities considered. On this line the deciles are marked and joined by straight lines to the median for the previous year.

**77. Average deviation.**—The average deviation (A.D.) is the arithmetic average of the deviations, all considered positive, from the mode, median, or arithmetic mean.

Theorem: *The average deviation is smallest when taken about the median.*

*Proof:* This is theorem IV of chapter VII (§65) with the word *deviation* replaced by the words *average deviation.*

Because of this theorem, it is theoretically best to compute the average deviation from the median. As a matter of practice there is little difference unless the median and arithmetic average differ widely.

The algebraic formula for the average deviation from the median is

$$\text{A.D.}_{Md} = \frac{\Sigma f|X - Md|^*}{\Sigma f} = \frac{\text{sum of deviations all considered positive}}{\text{number of deviations}}$$

with corresponding formulas for the average deviation from the mode, or arithmetic average.

$Md$ = median

$f$ = frequency in each class interval.

$X$ = mid-point of the class interval.

For a symmetrical (normal) distribution the average deviation is the abscissa of the center of gravity of the area under the right-hand half of the frequency curve.

We make the computation for the following data taken from Bulletin 110, Part 1, Bureau of Animal Husbandry, United States Department of Agriculture on "A Biometrical Study of Egg Production in the Domestic Fowl."

* These vertical bars are a direction to use all deviations as though they were positive, neglecting negative signs.

**Table 44.—Computation of Average Deviation from the Median for Frequency Distribution of Annual Egg Production in the Domestic Fowl.**

| *Annual Egg Production 1905–06* | *Mid-point* $X$ | *Frequency* $f$ | $X - Md$ | $f\|X - Md\|$ |
|---|---|---|---|---|
| 30– 44 | 37 | 1 | −105.5 | 105.5 |
| 45– 59 | 52 | 2 | − 90.5 | 181.0 |
| 60– 74 | 67 | 4 | − 75.5 | 302.0 |
| 75– 89 | 82 | 9 | − 60.5 | 544.5 |
| 90–104 | 97 | 13 | − 45.5 | 591.5 |
| 105–119 | 112 | 25 | − 30.5 | 762.5 |
| 120–134 | 127 | 24 | − 15.5 | 372.0 |
| 135–149 | 142 | 22 | − 0.5 | 11.0 |
| 150–164 | 157 | 32 | 14.5 | 464.0 |
| 165–179 | 172 | 17 | 29.5 | 501.5 |
| 180–194 | 187 | 20 | 44.5 | 890.0 |
| 195–209 | 202 | 9 | 59.5 | 535.5 |
| | | $\Sigma f = 178$ | $\Sigma f\|X - Md\| = 5{,}261.0$ | |

$$Md = 135 + \tfrac{11}{22} \times 15 = 142.5; \quad A.D._{Md} = \frac{5{,}261}{178} = 29.5$$

Fig. 37 illustrates table 44.

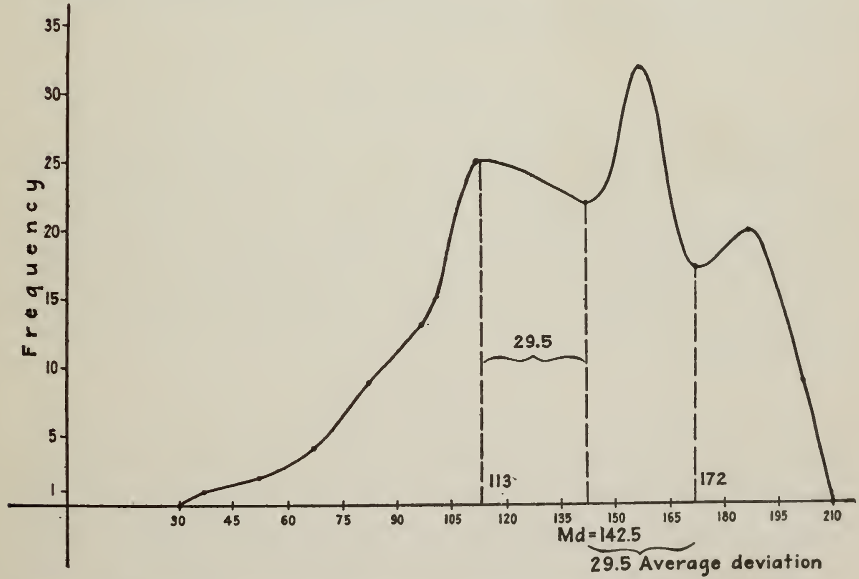


Fig. 37.—A Biometrical Study of Egg Production in the Domestic Fowl.

A deviation of 29.5 above the median gives 172. A deviation of 29.5 below the median gives 113. Thus, if 89 (or one-half) of the hens had laid exactly 113 eggs and the other half had laid exactly 172 eggs each, then the median would remain the same and the average deviation of the flock from this median would remain 29.5.

**78. Standard deviation.**—In computing the average deviation, it is found that some of the deviations are positive, some negative. The negative sign is ignored. The deviations are added, treating all deviations as positive. In order to avoid the necessity of ignoring the negative sign, each deviation may be squared before adding. Divide the sum by the number of deviations. Take the square root of this quotient. The result is called the *standard deviation.* The result is sometimes spoken of as the root-mean-square.

The computation of the standard deviation differs from that for the average deviation only in that the deviations are squared before multiplying by the frequencies. In computing the standard deviation, the deviations are always counted from the arithmetic average.

From the definition given above, the algebraic formula for the standard deviation (S.D. or $\sigma$) is

$$\sigma = \sqrt{\frac{\Sigma f(X - \overline{X})^2}{N}}$$

*Short-cut method.*—Whatever the mid-point of the class intervals, the deviations $X - \overline{X}$ are not usually integers. For example, the arithmetic average for the data on egg production, used in illustrating the computation of an average deviation, is 139.8. This leads to the squaring of such numbers as

$$157 - 139.8 = 17.2$$
$$172 - 139.8 = 32.2$$

This is not an unusually difficult feat. But any method of procedure is considered difficult when compared with another method which accomplishes the same purpose with less work.

What we do is to choose an arbitrary origin $X_o$ at the mid-point of a class interval near the arithmetic average.

Compute the deviations $X - X_o$ from this arbitrary origin. We have

$$X - X_o = (X - \bar{X}) + (\bar{X} - X_o)$$

Multiply each squared deviation by its frequency $f$. Add together all such products and divide the result by the total frequency $N$. We have

$$\frac{\Sigma f(X - X_o)^2}{N} = \frac{\Sigma f(X - \bar{X})^2}{N} + 2(\bar{X} - X_o)\frac{\Sigma f(X - \bar{X})}{N} + \frac{\Sigma f(\bar{X} - X_o)^2}{N}$$

But

$$\frac{\Sigma f(X - \bar{X})}{N} = 0 \text{ and } \frac{\Sigma f(X - \bar{X})^2}{N} = \sigma^2$$

Whence

$$\sigma = \sqrt{\frac{\Sigma f(X - X_o)^2}{N} - (\bar{X} - X_o)^2}$$

or[2]

$$\sigma^2 = \sigma_o^2 - d^2$$

Thus we arrive at the rule:

1. Compute the deviations from a convenient arbitrary origin.
2. Square these deviations.
3. Multiply each squared deviation by its frequency.
4. Add.
5. Divide by the total frequency.
6. Find the difference between the arbitrary origin and the arithmetic average.
7. Square the number found in step 6.
8. Subtract the result found in 7 from the result found in 5.
9. Extract the square root of the result found in 8.

In table 45 is given the form for the computation for the data on egg production given in table 44.

[2] We have put $\sigma_o^2 = \frac{\Sigma f(X - X_o)^2}{N}$. $d^2 = (\bar{X} - X_o)^2$.

**Table 45.—Computation of Standard Deviation for Frequency Distribution of Annual Egg Production in the Domestic Fowl.**

| $X$ | $f$ | $X - X_o$ | $f(X - X_0)$ | $f(X - X_o)^2$ | $f(\overline{X - X_o + 1})^2$ |
|---|---|---|---|---|---|
| 37 | 1 | −90 | − 90 | 8,100 | 7,921 |
| 52 | 2 | −75 | −150 | 11,250 | 10,952 |
| 67 | 4 | −60 | −240 | 14,400 | 13,924 |
| 82 | 9 | −45 | −405 | 18,225 | 17,424 |
| 97 | 13 | −30 | −390 | 11,700 | 10,933 |
| 112 | 25 | −15 | −375 | 5,625 | 4,900 |
| 127 | 24 | 0 | 0 | 0 | 24 |
| 142 | 22 | 15 | 330 | 4,950 | 5,632 |
| 157 | 32 | 30 | 960 | 28,800 | 30,752 |
| 172 | 17 | 45 | 765 | 34,425 | 35,972 |
| 187 | 20 | 60 | 1,200 | 72,000 | 74,420 |
| 202 | 9 | 75 | 675 | 50,625 | 51,984 |
|  | 178 | ..... | 2,280 | 260,100 | 264,838 |

$$\overline{X} = 127 + \frac{2{,}280}{178} = 139.8; \sigma = \sqrt{\frac{260{,}100}{178} - (12.8)^2} = 36.02$$

Charlier check:[3] 264,838 = 260,100 + 2(2,280) + 178.

A still further reduction in the labor involved is obtained by using the step deviation process[4] whenever the class interval is not unity. This is illustrated by the computations shown in table 46.

[3] By computing an additional column one obtains a check upon the accuracy of the footings of the other columns. We have, by simple algebra,

$$\Sigma f(\overline{X - X_o + 1})^2 = \Sigma f(X - X_o)^2 + 2\Sigma f(X - X_o) + \Sigma f$$

This is known as the Charlier check.

[4] In the past, economists have shown a preference for the average deviation. Biologists have always shown a preference for the standard deviation. The standard deviation, from a theoretical point of view, is to be preferred. It gives more weight to the extreme deviations. It is used (see Chapter IX) in finding equations of lines of regression, and in determining the Pearsonian coefficient of correlation, and (see chapter XV) in fitting a normal probability curve to frequency data and in the determination of probable errors.

Table 46.—Step Deviation Process for Computation of Standard Deviation for Frequency Distribution in Annual Egg Production.

| $X$ | $f$ | *Step* $s$ | $fs$ | $fs^2$ | $f(s+1)^2$ |
|---|---|---|---|---|---|
| 37 | 1 | −6 | − 6 | 36 | 25 |
| 52 | 2 | −5 | −10 | 50 | 32 |
| 67 | 4 | −4 | −16 | 64 | 36 |
| 82 | 9 | −3 | −27 | 81 | 36 |
| 97 | 13 | −2 | −26 | 52 | 13 |
| 112 | 25 | −1 | −25 | 25 | 0 |
| 127 | 24 | 0 | 0 | 0 | 24 |
| 142 | 22 | 1 | 22 | 22 | 88 |
| 157 | 32 | 2 | 64 | 128 | 288 |
| 172 | 17 | 3 | 51 | 153 | 272 |
| 187 | 20 | 4 | 80 | 320 | 500 |
| 202 | 9 | 5 | 45 | 225 | 324 |
|  | 178 | . . . | 152 | 1,156 | 1,638 |

$$\overline{X} = 127 + \tfrac{152}{178} \times 15 = 139.8; \ \sigma = \sqrt{\frac{1{,}156}{178}(15)^2 - (12.8)^2} = 36.02$$

Charlier check: 1,638 = 1,156 + 2(152) + 178.

The formula

$$\sigma^2 = \sigma_o{}^2 - d^2$$

for the computation of the standard deviation by making use of an arbitrary origin can be kept in mind by remembering the following geometrical diagram:

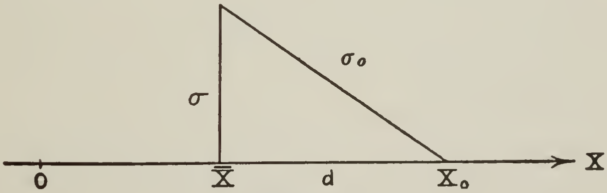


Fig. 38.

**79. Coefficient of dispersion.**—The significance to be attached to a value for a measure of dispersion depends upon the size of the corresponding average from which it is computed. Thus, a variation of 1 lb. from an average weight of 8 lbs. for turkeys is more important than a variation of 1 lb. from an average weight of 800 lbs. for cattle. A daily variation of 2 qts. in the milk production of a cow is more important than a monthly variation of 2 qts. A standard deviation of one dollar per day among the employees of one factory represents a greater variation in wages than a standard deviation of one dollar per month among the employees of another factory. On the other hand, a standard deviation of one inch in the measurements of the lengths of feet having an average length of ten inches is of the same relative significance as a standard deviation of seven inches in the measurements of the heights of a group of individuals having an average height of 70 inches.

Thus, it is seen that, in order to obtain a significant measure of dispersion for the purpose of comparing two distributions, it is desirable to take account of the relative size of the items. This is done by dividing the measure of dispersion by the average from which it is computed. The resulting number is called a *coefficient of dispersion.* This number is usually quite small. Multiplied by 100—expressed as a percentage—the result is termed a *coefficient of variability.*

In the case of the quartile deviation, a coefficient is obtained by using the following formula:

$$\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{Q_3 - Q_1}{2} \div \text{median}$$

**80. Lorenz curve.**—A Lorenz[5] curve is a cumulative frequency graph which shows graphically the manner in which the distribution of wealth, income, wages, etc., departs from equal distribution.

Let us illustrate by using the data for the number and value of estates of men dying in Massachusetts, period 1889–1891, taken from W. I. King, "Wealth and Income of the Population of the United States," p. 71. Under an equal distribution of wealth, 25 per cent of the people should possess 25 per cent of the wealth, 50 per cent of the people should possess 50 per cent of the wealth, and so on. This equal distribution would be exhibited by the straight line *AB* in fig. 39.

[5] The reader should consult the previous discussion in §48.

**Table 47.—Number and Value of Estates of Men Dying in Massachusetts.**

| *Value of Estate in Thousands of Dollars* | *Per Cent of Total Number* | *Per Cent of Total Value* | *Average Value Estate in Class* |
|---|---|---|---|
| Under 0.5 | 65.864 | 4.568 | 375 |
| 1 | 70.036 | 5.248 | 716 |
| 5 | 86.298 | 12.862 | 1,733 |
| 10 | 92.002 | 20.082 | 3,931 |
| 25 | 96.588 | 32.807 | 5,620 |
| 50 | 98.111 | 42.131 | 8,226 |
| 100 | 99.066 | 53.716 | 11,940 |
| 200 | 99.567 | 65.629 | 16,900 |
| 300 | 99.735 | 72.847 | 33,812 |
| 400 | 99.815 | 77.530 | 69,583 |
| 500 | 99.866 | 81.610 | 218,220 |
| All estates | 100.000 | 100.000 | 622,000 |

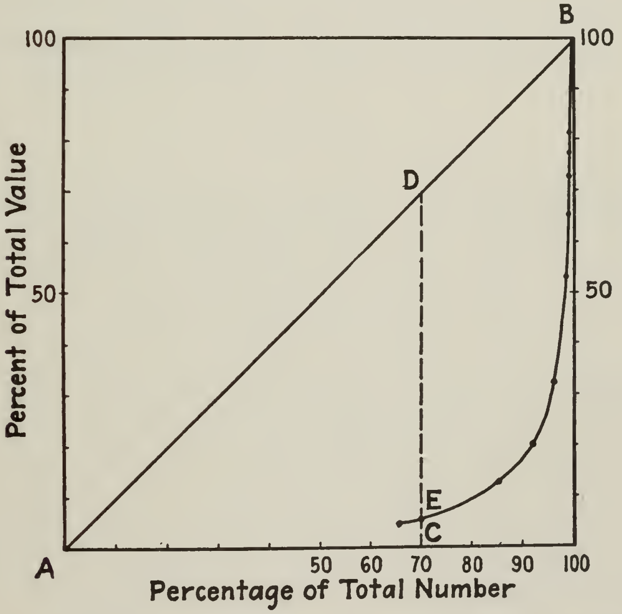


**Fig. 39.**—Number and Value of Estates of Men Dying in Massachusetts, 1889–1891.

Under an equal distribution, 70 per cent of the people should possess 70 per cent of the wealth, as shown by the line *CD* in fig. 39, instead of approximately 5 per cent as shown by the line *CE*. The greater the inequality in the distribution, the greater is the departure of the cumulative graph from the straight line.

**81. Skewness.**—Skewness is a lack of symmetry. For an asymmetric distribution the range is usually greater on one side of the mode than on the other. The curve is said to have a longer tail on one side than on the other. The existence of skewness implies some predominant cause which influences the selection. Thus, a curve of distribution of wealth would have a long tail to the right. There is no upper limit to the amount of wealth one may have, and a few people do possess enormous wealth.

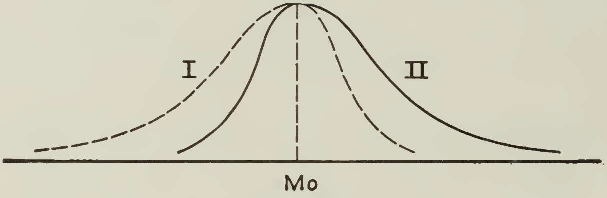


Fig. 40.

Two curves may have the same mode and the same measure of dispersion but still represent markedly different distribution. This is illustrated by the two curves in fig. 40. Skewness is *positive* when the longer range is to the right of the mode, as in curve II. Skewness is *negative* when the longer range is to the left of the mode, as in curve I.

A distribution is generally symmetrical if it represents many measurements of the same item, for positive and negative errors are equally likely to occur. Whenever the data represent the measurement of the same characteristic for many different items, skewness is likely to be present. Frequency distributions for economic data are usually skewed.

Several formulas have been devised for measuring skewness. We will give three:

1. Pearson's.
2. Quartile.
3. Third moment.

**82. Pearson's measure of skewness.**—It has been noted in the chapter on averages that, for a non-symmetric distribution, the mode, median, and arithmetic average are separated. The greater the departure from symmetry, the greater the separation. Since skewness is a measure of the shape of the curve and not of its size, it seems desirable to use a ratio as a coefficient of the measure of skewness.

Karl Pearson devised the following coefficient of the measure of skewness. It is perhaps the most generally used.

$$\text{Skewness} = \frac{\text{Arithmetic mean} - \text{Mode}}{\text{Standard deviation}}$$

The mode is not always well defined. For such cases some other measure of skewness must be used.

**83. Quartile measure of skewness.**—If the curve is asymmetric, the range from the first quartile to the median is not the same as the range from the median to the third quartile. Thus, the difference

$$(Q_3 - Md) - (Md - Q_1)$$

is a measure of skewness. In order to reduce this to a relative number, it is customary to divide by $Q_3 - Q_1$. Thus, we have for a coefficient, sk., of skewness

$$\text{sk.} = \frac{(Q_3 - \text{Md}) - (\text{Md} - Q_1)}{Q_3 - Q_1}$$

**84. Third moment.**—The sum of the deviations from the arithmetic average is zero regardless of the form of the distribution. Squaring the deviations from the arithmetic average destroys the distinction between positive and negative deviations. The cubes of these deviations maintain the distinction between positive and negative deviations. Hence, we have a measure of skewness by taking the cube root of the arithmetic average of the sum of the cubes of the deviations. A coefficient of skewness is obtained by dividing by the standard deviation.

$$\text{sk.} = \frac{\sqrt[3]{\dfrac{\Sigma f(X - X)^3}{N}}}{\sigma}$$

## Chapter IX

# CORRELATION

**85. Meaning of correlation.**—It has been observed that, in some instances, there exists a relationship between pairs of variables. Thus, there is a relation between age and height of individuals. There is a normal weight for an individual of given height. Growing children are watched carefully and weighed regularly by their parents and by the school authorities to determine whether they have the proper weight for their respective heights. This relation between height and weight is *not a precise proportionality.* The weight assigned to a given height is an average of the weights of many individuals of different weights. Thus, for a boy whose height is 58 inches, one table gives the following distribution of weight in pounds for the age given:

| *Age:* | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|
| *Lbs.:* | 84 | 85 | 86 | 87 | 88 | 89 | 90 |

There are times when the existence of a relationship between pairs of variables enables us to predict with a fair degree of accuracy what will happen in the future. Thus, if the production of corn is increased, we may expect the price, other factors remaining constant, to decrease and to remain down for several months. If rainfall is below normal, we expect crop production to be below normal. If the price of wheat increases, we look to see the price of flour increase. Likewise, if the wholesale price of flour increases, we may expect an increase in the retail price of bread. These increases are not, however, precisely proportional to the prices of wheat and of flour. The relation between the annual wholesale price quotations of flour and retail prices of bread in cents per pound at Minneapolis, Kansas City, and Boston for the decade 1913–1923 is shown in the following table: (*Source: Brown, Edmund, "Marketing," p.* 58).

**Table 48.—Annual Wholesale and Retail Prices of Flour and Bread in the Cities Given.**

| *Year* | Minneapolis | | Kansas City | | Boston | |
|---|---|---|---|---|---|---|
| | *Flour Wholesale* | *Bread Retail* | *Flour Wholesale* | *Bread Retail* | *Flour Wholesale* | *Bread Retail* |
| 1913 | 2.2 | 5.6 | 2.1 | 6.0 | 2.5 | 5.9 |
| 1914 | 2.4 | 5.7 | 2.2 | 6.1 | 2.9 | 6.0 |
| 15 | 3.3 | 6.4 | 3.0 | 7.1 | 3.7 | 6.5 |
| 16 | 3.6 | 6.9 | 3.2 | 7.8 | 4.1 | 6.8 |
| 17 | 5.7 | 9.5 | 5.6 | 10.2 | 6.2 | 8.8 |
| 18 | 5.2 | 8.9 | 5.2 | 10.0 | 5.6 | 9.0 |
| 19 | 6.2 | 9.5 | 5.9 | 9.9 | 6.7 | 9.4 |
| 20 | 6.7 | 10.7 | 6.4 | 12.5 | 7.2 | 11.2 |
| 21 | 4.5 | 9.5 | 4.0 | 10.4 | 5.0 | 10.0 |
| 22 | 3.9 | 8.8 | 3.5 | 7.8 | 4.4 | 8.5 |
| 1923 | 3.4 | 9.0 | 3.1 | 8.2 | 3.9 | 8.4 |

Whenever two variables are so related that a change in one is accompanied by a change in the other in such a way that an increase in the one is accompanied by an increase in the other, or a decrease in the one by a decrease in the other, and the greater the magnitude of the change in the one, the greater the amount of the change in the other, then the variables are said to be correlated.

If, as one variable increases, there is a general tendency for the other variable to increase, correlation is said to be *positive.* If with an increase in the one variable there is a general tendency for the other variable to decrease, correlation is said to be negative.

**86. Dot diagram.**—As a first step in obtaining a measure of the relationship between a pair of variables which are correlated we consider a dot diagram.

Let us measure the height of oats plants in centimeters and the average number of kernels per culm per plant. We obtain thus a set of number pairs $(x_1, y_1)$, $(x_2, y_2)$, . . . , $(x_n, y_n)$. Such a set of number pairs is obtained when one measures the height and weight of individuals, the length and weight of ears of corn, the weight in ounces after death of the heart and kidneys of healthy males.

These number pairs may be used as the coördinates of points in a plane. The map made by plotting the points

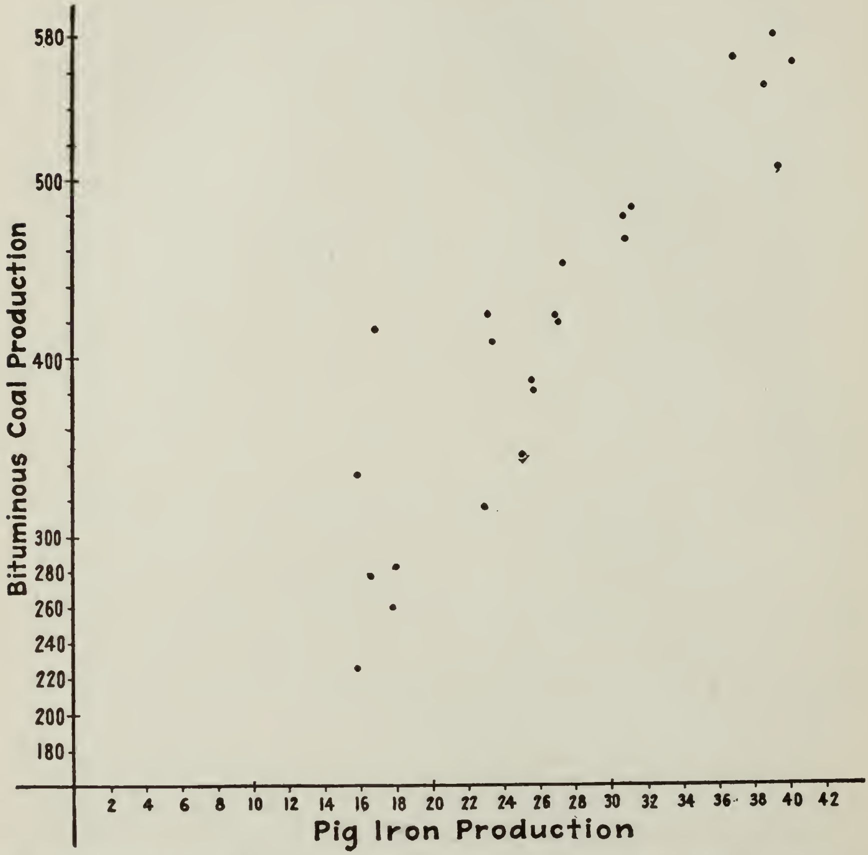


Fig. 41.—Dot Diagram Illustrating the Correlation between Pig Iron Production and Bituminous Coal Production.

corresponding to each number pair is called a *dot diagram* or a *scatter diagram.* In general, it will be observed that for each value of one variable there are many values of the other variable.

Whenever correlation exists, it will be found that the plotted points tend to concentrate in a band of greater or less width. In general, the narrower this band, the greater the degree of correlation.

A dot diagram is shown in fig. 41. The data for this diagram are given in table 49. For any given year the production of pig iron is used as an abscissa and the production of bituminous coal as an ordinate.

**Table 49.—Pig Iron Production in Millions of Long Tons: Bituminous Coal Production in Long Tons (2,000 Lbs.)**

| *Year* | *Pig Iron* | *Bituminous Coal* | *Year* | *Pig Iron* | *Bituminous Coal* |
|---|---|---|---|---|---|
| 1901 | 15.9 | 225.8 | 1913 | 31.0 | 478.4 |
| 02 | 17.8 | 260.2 | 14 | 23.3 | 422.7 |
| 03 | 18.0 | 282.7 | 15 | 29.9 | 442.6 |
| 04 | 16.5 | 278.7 | 16 | 39.4 | 502.5 |
| 05 | 23.0 | 315.0 | 17 | 38.6 | 551.8 |
| 1906 | 25.3 | 342.9 | 1918 | 39.1 | 579.4 |
| 07 | 25.8 | 394.8 | 19 | 31.0 | 465.9 |
| 08 | 15.9 | 332.6 | 20 | 36.9 | 568.7 |
| 09 | 25.8 | 379.7 | 21 | 16.7 | 415.9 |
| 10 | 27.3 | 417.1 | 22 | 27.2 | 422.3 |
| 1911 | 23.6 | 405.9 | 1923 | 40.4 | 564.2 |
| 12 | 29.7 | 450.1 | 1924 | 31.4 | 483.3 |

*Statistical Abstract of U. S., 1924, pp. 697, 704.*

**87. Correlation table.**—A correlation table is a table of double entry for pairs of measured characters. When the number of such measurements becomes large, the table becomes unwieldy. In order to condense such a table, some class interval is adopted for the measurements of each variate. Then for each interval in the measurements of one variate an ordinary frequency table is formed for the measurements of the other variate.

A correlation table may be obtained from the dot diagram by subdividing the coördinate area into equal rectangular compartments and then writing within each compartment the number of dots which fall within it.

The following is a correlation table of monthly pig iron tonnage and the monthly price index:

Table 50.—Correlation Table for the Monthly Price and Production of Pig Iron.

| Monthly Pig Iron Tonnage (*in Millions of Tons*) | *Monthly Price Index*<br>R1 - Rv | | | | | *Arith. Mean* | Totals |
|---|---|---|---|---|---|---|---|
| | 7.50 to 8.00 | 8.00 to 8.50 | 8.50 to 9.00 | 9.00 to 9.50 | 9.50 to 10.00 | | |
| 28 to 32........... | ..... | ..... | ..... | 4 | 4 | 9.50 | 8 |
| 24 to 28........... | ..... | ..... | 4 | 16 | 1 | 9.18 | 21 |
| 20 to 24........... | ..... | 9 | 26 | 7 | 5 | 8.84 | 47 |
| 16 to 20........... | 8 | 21 | 14 | 3 | 1 | 8.41 | 47 |
| 12 to 16........... | 10 | 8 | 2 | 2 | 1 | 8.23 | 23 |
| 8 to 12........... | 7 | 3 | ..... | ..... | .. | 7.90 | 10 |
| *Arith. Mean*....... | 14.16 | 17.51 | 20.78 | 24.12 | 24 | | |
| *Totals*........... | 25 | 41 | 46 | 32 | 12 | | 156 |

*A. R. Crathorne, "Calculation of the Correlation Ratio," Journal of American Statistical Association, Sept., 1922, pp. 394–396.*

Each number of the table occupies what is termed a *cell* of the table. The numbers in a column are termed an $x$ array of $y$'s. The numbers in any given row are termed a $y$ array of $x$'s.

Compute the arithmetic average tonnage for the distribution given in each column. We give here the computation for the first column:

$$\frac{8 \times 18 + 10 \times 14 + 7 \times 10}{8 + 10 + 7} = 14.16$$

Compute the arithmetic average price index for the distribution given in each row and tabulate as indicated. We give here the computation for the second row:

$$\frac{4 \times 8.75 + 16 \times 9.25 + 1 \times 9.75}{4 + 16 + 1} = 9.18$$

It is obvious from these data that there is some relation between tonnage and price index of pig iron. It is our aim to find a measure of this relationship.

**88. Regression lines.**—Let us plot as ordinates the means of the columns and as abscissa the corresponding price index. These points are approximately on a straight line $C_1C_2$, as shown in fig. 42. This line is called the *line of regression* of mean tonnage with respect to price index. Generally this line is termed the line of regression of $y$ on $x$.

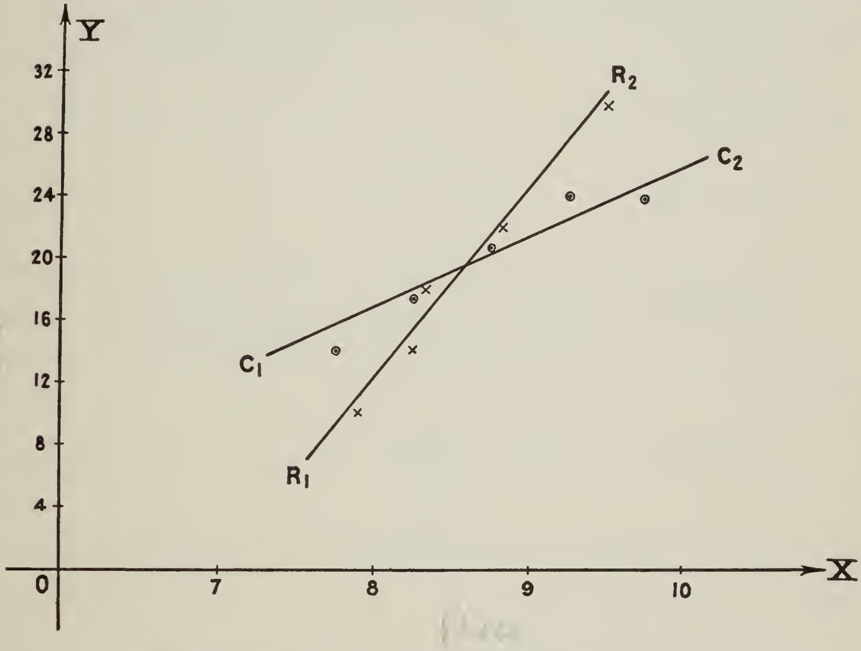


Fig. 42.—Lines of Regression for the Data in Table 50.

Let us plot as abscissa the means of the rows, and as ordinates the corresponding tonnage. These points are approximately on a straight line $R_1R_2$. This line is called the line of regression of mean-price index with respect to monthly tonnage of pig iron. Generally this line is termed the line of regression of $x$ on $y$.

Whenever, as in the present instance, the plotted points are approximately on a straight line, the regression is said to be *linear*. In some instances the points so plotted do not fall along a straight line. The regression is then said to be *non-linear*.

Whenever, as in the present illustration, the means of the $y$ arrays are approximately on a straight line, we assume a linear relation of the form

$$Y = mX + B^* \tag{1}$$

and proceed to determine $m$ and $B$ in such a way that the sum of the zero-th and first moments of the ordinates of all of the points of the dot diagram shall be the same as that for the ordinates as computed from the assumed equation.

Let us take as a new $y$-axis a vertical line through the mean of the $X$'s; then

$$x = X - \overline{X}$$

The equation (1) now becomes[1]

$$Y = mx + b$$

Referred to this new set of axes, let the points of the dot diagram which we desire to fit by a straight line be $(x_1, Y_1)$, $(x_2, Y_2)$, . . . , $(x_n, Y_n)$. We have then for the zero-th moment:

$$\sum_1^n{}_i Y_i = \sum_1^n{}_i (mx_i + b)$$

$$= m\sum_1^n x_i + nb$$

But

$$\Sigma x_i = 0$$

Hence

$$b = \frac{\Sigma Y_i}{n}$$

Computing first moments, we find

$$\begin{aligned} \Sigma x_i Y_i &= \Sigma x_i(mx_i + b) \\ &= m\Sigma x_i^2 + b\Sigma x_i \end{aligned}$$

* $B$ is the $y$ coördinate of the point where the line crosses the $y$-axis. $m$ indicates the slope (steepness) of the line. The greater $m$, the steeper the line.

[1] Substituting in (1) we have

$$\begin{aligned} Y &= m(x - \overline{X}) + B \\ &= mx + (B - m\overline{X}) \\ &= mx + b \end{aligned}$$

if we put

$$b = B - m\overline{X}\cdot$$

Whence

$$m = \frac{\Sigma x_i Y_i}{\Sigma x_i^2}$$

Thus, we have for the equation of the line of regression of $Y$ on $x$

$$Y = \frac{\Sigma x_i Y_i}{\Sigma x_i^2} x + \frac{\Sigma Y_i}{n}$$

Let us take as a new $x$-axis a horizontal line through the mean of the $Y$'s; then

$$y = Y - \overline{Y}$$

The equation of the line of regression[2] of $y$ on $x$ is then

$$\mathbf{y} = \frac{\Sigma \mathbf{x}_i \mathbf{y}_i}{\Sigma \mathbf{x}_i^2} \mathbf{x}$$

$$= \frac{\Sigma \mathbf{x}_i \mathbf{y}_i}{\mathbf{n} \sigma_x^2} \mathbf{x}$$

If we *define* a number $r$ as follows:

$$r = \frac{\Sigma x_i y_i}{n \sigma_x \sigma_y}$$

the equation of the line of regression of $y$ on $x$ becomes

(2) $$y = r \frac{\sigma_y}{\sigma_x} x$$

In like manner, we can show that the equation of the line of regression of $x$ on $y$ is

(3) $$x = r \frac{\sigma_x}{\sigma_y} y$$

[2] $$Y = \frac{\Sigma x_i Y_i}{\Sigma x_i^2} x + \frac{\Sigma Y_i}{n}$$

Whence

$$y + \overline{Y} = \frac{\Sigma x_i (y_i + \overline{Y})}{\Sigma x_i^2} x + \frac{\Sigma Y_i}{n},$$

which reduces to

$$y = \frac{\Sigma x_i y_i}{\Sigma x_i^2} x$$

since

$$\overline{Y} = \frac{\Sigma Y_i}{n}$$

and

$$\frac{\Sigma x_i \overline{Y}}{\Sigma x_i^2} x = \frac{x \overline{Y}}{\Sigma x_i^2} \cdot \Sigma x_i = 0,$$

for

$$\Sigma x_i = 0.$$

**89. Meaning of regression.**—From the computations in §90, the line of regression of $x$ on $y$ is

$$x = 0.08y$$

For this equation the origin is at the arithmetic mean value of monthly tonnage and monthly price index. Let us change the origin so that the variables represent actual tonnage and actual price index. We have as equations of transformation, since $\overline{X} = 8.64$ and $\overline{Y} = 19.795$,

$$X = x + 8.64 \text{ and } Y = y + 19.795$$

Our line of regression of mean price index with respect to monthly tonnage now becomes

$$X - 8.64 = 0.08(Y - 19.80)$$

or

$$X = 0.08Y + 7.06$$

From this equation we compute that, for a pig iron tonnage of $y = 22$ millions of tons, we shall have an average of

$$X = 0.08 \times 22 + 7.06 = 8.83$$

for the monthly price index. Consulting the table 50, we find that, for the sample under consideration, the average monthly price index was 8.84 for all price indices which were in the class interval 20 to 24 with respect to production. Thus, we see that, given a value of one variable, the equations of regression enable us to make a fair estimate of the average value of the other variable.

**90. Pearson's product moment correlation coefficient.**—The number $r$ defined above is Pearson's product moment coefficient of correlation. It is easy to verify from the above definition of $r$, if we make use of the definitions[3] of $\sigma_x$ and $\sigma_y$, that

$$r = 1 - \frac{1}{2n}\sum\left(\frac{x}{\sigma_x} - \frac{y}{\sigma_y}\right)^2$$

$$= -1 + \frac{1}{2n}\sum\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right)^2$$

From these two equations we see that $r$ cannot be greater than $+1$ nor less than $-1$.

$$r = +1, \text{ when } x \cdot \sigma_y = y \cdot \sigma_x$$
$$r = -1, \text{ when } x \cdot \sigma_y = -y \cdot \sigma_x$$

[3] $\sigma_x^2 = \frac{\Sigma x_i^2}{n}$; $\sigma_y^2 = \frac{\Sigma y_i^2}{n}$

In either case there is said to be perfect correlation between the two sets of quantities.

For purposes of computation it is desirable to express $r$ differently. Indicate by $X$ and $Y$ new variables referred to any convenient point as origin. Then

$$r = \frac{\Sigma x_i y_i}{n\sigma_x\sigma_y} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2}\sqrt{\Sigma(Y_i - \overline{Y})^2}}$$

$$= \frac{\frac{1}{n}\Sigma(X_iY_i) - \overline{X}\,\overline{Y}}{\sqrt{\frac{1}{n}\Sigma X_i^2 - \overline{X}^2}\ \sqrt{\frac{1}{n}\Sigma Y_i^2 - \overline{Y}^2}}$$

In order to carry out the computations indicated by this formula, certain forms for carrying on the work are desirable. These are indicated in table 51.

The number 16 in the column $XY$ is obtained as follows: $Y = 4$. Each frequency in the corresponding row is multiplied by the proper value of $X$. Thus

$$4[5(1.0) + 7(0.5) + 26(0) + 9(-0.5)] = 16$$

**Table 51.—Computation of Correlation of Coefficient $r$ for Monthly Pig Iron Tonnage and Monthly Price Index of Pig Iron.**

| Monthly Tonnage | *Monthly Price Index* | | | | | $f_y$ | $Y$ | $f_y \cdot Y$ | $f_y \cdot Y^2$ | $XY$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 7.50<br>8.00 | 8.00<br>8.50 | 8.50<br>9.00 | 9.00<br>9.50 | 9.50<br>10.00 | | | | | |
| 28 to 32 | ...... | ....... | .. | 4 | 4 | 8 | 12 | 96 | 1,152 | 72 |
| 24 to 28 | ...... | ....... | 4 | 16 | 1 | 21 | 8 | 168 | 1,344 | 72 |
| 20 to 24 | ...... | 9 | 26 | 7 | 5 | 47 | 4 | 188 | 752 | 16 |
| 16 to 20 | 8 | 21 | 14 | 3 | 1 | 47 | 0 | 0 | 0 | 0 |
| 12 to 16 | 10 | 8 | 2 | 2 | 1 | 23 | −4 | −92 | 368 | 48 |
| 8 to 12 | 7 | 3 | .. | .... | .... | 10 | −8 | −80 | 640 | 68 |
| $f_x$ | 25 | 41 | 46 | 32 | 12 | 156 | .. | 280 | 4,256 | 276 |
| $X$ | − 1.0 | − 0.5 | 0 | 0.5 | 1.0 | | $\overline{Y} = \frac{280}{156} = 1.795$ | | | |
| $f_x \cdot X$ | −25 | −20.5 | 0 | 16 | 12 | −17.5 | $\overline{X} = \frac{-17.5}{156} = -0.112$ | | | |
| $f_x \cdot X^2$ | 25 | 10.25 | 0 | 8 | 12 | 55.25 | $\sigma_y^2 = \frac{4{,}256}{156} - \overline{Y}^2 = 24.06$ | | | |
| $XY$ | 96 | 10 | 0 | 98 | 72 | 276 | $\sigma_y = 4.9$ | | | |

$$\sigma_x^2 = \frac{55.25}{156} - \overline{X}^2 = 0.342; \sigma_x = 0.59$$

$$\frac{1}{n}\Sigma XY = \tfrac{276}{156} = 1.77$$

$$r = \frac{\frac{1}{n}\Sigma XY - \overline{XY}}{\sigma_x \cdot \sigma_y} = \frac{1.77 + 0.201}{2.89} = 0.68$$

$$y = r\frac{\sigma_y}{\sigma_x}x = 5.8x;\ Y = 5.8X - 30.31$$

$$x = r\frac{\sigma_x}{\sigma_y}y = 0.08y; X = 0.08Y + 7.06$$

## Exercises.

For each of the following tables compute Pearson's product moment correlation coefficient:

1. Correlation of number of culms per oat plant and total yield of plant in grams (*Source: Love-Leighty*).

$r = 0.712$

| *Yield* | *Number of Culms per Plant* | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| 0–1 | 3 | | | | | |
| 1–2 | 28 | 19 | 3 | | | |
| 2–3 | 18 | 66 | 20 | 1 | .. | 1 |
| 3–4 | 1 | 42 | 58 | 7 | 1 | |
| 4–5 | .. | 7 | 59 | 11 | 3 | |
| 5–6 | .. | .. | 26 | 14 | 2 | |
| 6–7 | .. | .. | .. | 4 | 3 | |
| 7–8 | .. | .. | 1 | 1 | | |
| 8–9 | .. | .. | .. | .. | 1 | |

2. Correlation of heights of adult children and parents. Data for children of 205 mid-parents of various statures (*Source: Galton-Davenport*).

| *Heights* of Mid-parents* | *Heights of Adult Children in Inches* | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Above* | *73.2* | *72.2* | *71.2* | *70.2* | *69.2* | *68.2* | *67.2* | *66.2* | *65.2* | *64.2* | *63.2* | *62.2* | *Below* |
| *Above* | | 3 | 1 | | | | | | | | | | | |
| 72.5 | 4 | 2 | 7 | 2 | 1 | 2 | 1 | | | | | | | |
| 71.5 | 2 | 2 | 9 | 4 | 10 | 5 | 3 | 4 | 3 | 1 | | | | |
| 70.5 | 3 | 3 | 4 | 7 | 14 | 18 | 12 | 3 | 1 | 1 | .. | 1 | .. | 1 |
| 69.5 | 5 | 4 | 11 | 20 | 25 | 33 | 20 | 27 | 17 | 4 | 16 | 1 | | |
| 68.5 | .. | 3 | 4 | 18 | 21 | 48 | 34 | 31 | 25 | 16 | 11 | 7 | | 1 |
| 67.5 | .. | .. | 4 | 11 | 19 | 38 | 28 | 38 | 36 | 15 | 14 | 5 | 3 | |
| 66.5 | .. | .. | .. | .. | 4 | 13 | 14 | 17 | 17 | 2 | 5 | 3 | 3 | |
| 65.5 | .. | .. | 1 | 2 | 5 | 7 | 7 | 11 | 11 | 7 | 5 | 9 | | |
| 64.5 | .. | .. | .. | .. | .. | .. | .. | 5 | 5 | 1 | 4 | 4 | 1 | 1 |
| *Below* | .. | .. | .. | .. | .. | 1 | 1 | 2 | 2 | 1 | 4 | 2 | .. | 1 |

* Height of mid-parent is the mean height of the two parents.

3. Height in inches and weights in pounds of Glasgow schoolboys, ages 4.5 to 5.5. (*Source: Biometrika, Vol. XI, p. 62.*)

$r = 0.72; \sigma_x = 0.90; \sigma_y = 0.83$

| *Height* | *Weight* | | | | | | |
|---|---|---|---|---|---|---|---|
| | 26 | 31 | 36 | 41 | 46 | 51 | *Totals* |
| 31 | 2 | .. | .. | .. | .. | .. | 2 |
| 34 | 5 | 15 | 5 | .. | .. | .. | 25 |
| 37 | 1 | 18 | 72 | 8 | .. | .. | 99 |
| 40 | .. | 5 | 87 | 90 | 7 | 1 | 190 |
| 43 | .. | .. | 4 | 35 | 21 | 5 | 65 |
| 46 | .. | .. | 1 | .. | 2 | .. | 3 |

4. Correlation table of stature with left cubit (forearm). From measurements on Cairo-born Egyptians. (*Source: Biometrika, Vol. XI, p. 81.*)

$r = 0.803$; stature $= 2.245\ 1(\text{cubit}) + 631.0.$

| | Groups | *Stature in cm.* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 148<br>151 | 152<br>155 | 156<br>159 | 160<br>163 | 164<br>167 | 168<br>171 | 172<br>175 | 176<br>179 | 180<br>183 | 184<br>187 | *Totals* |
| *Left Forearm in mm.* | 388–402 | 3 | .. | .. | .. | .. | .. | .. | .. | .. | .. | 3 |
| | 403–417 | 1 | 15 | 3 | 1 | .. | .. | .. | .. | .. | .. | 20 |
| | 418–432 | .. | 6 | 33 | 27 | 5 | 1 | .. | .. | .. | .. | 72 |
| | 433–447 | .. | 4 | 27 | 76 | 44 | 8 | 1 | .. | .. | .. | 160 |
| | 448–462 | .. | 3 | 9 | 60 | 91 | 54 | 7 | .. | .. | .. | 224 |
| | 463–477 | .. | .. | 1 | 17 | 54 | 77 | 29 | 6 | 2 | .. | 186 |
| | 478–492 | .. | .. | .. | 3 | 17 | 26 | 36 | 16 | 2 | .. | 100 |
| | 493–507 | .. | .. | .. | 1 | 1 | 1 | 12 | 6 | 5 | .. | 26 |
| | 508–522 | .. | .. | .. | .. | .. | .. | 2 | 3 | 5 | .. | 10 |
| | 523–537 | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | 0 |
| | 538–552 | .. | .. | .. | .. | .. | .. | .. | .. | .. | 1 | 1 |
| | *Totals* ....... | 4 | 28 | 73 | 185 | 212 | 167 | 87 | 31 | 14 | 1 | 802 |

5. Correlation table of stature with left foot. (*Source: Biometrika, Vol. XI, p. 81.*)

$r = 0.715$, stature (cm.) = 0.343 (feet in mm.) + 77.44 cm.

| | *Groups* | *Stature in cm.* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 148<br>151 | 152<br>155 | 156<br>159 | 160<br>163 | 164<br>167 | 168<br>171 | 172<br>175 | 176<br>179 | 180<br>183 | 184<br>187 | *Totals* |
| *Left Foot in mm.* | 222–228 | 1 | 1 | 1 | 2 | .. | .. | .. | .. | .. | .. | 5 |
| | 229–235 | 2 | 3 | 8 | 6 | .. | .. | .. | .. | .. | .. | 19 |
| | 236–242 | .. | 10 | 23 | 24 | 10 | 3 | .. | .. | .. | .. | 70 |
| | 243–249 | 1 | 9 | 16 | 41 | 35 | 8 | .. | 1 | .. | .. | 111 |
| | 250–256 | .. | 5 | 21 | 65 | 53 | 30 | 4 | .. | .. | .. | 178 |
| | 257–263 | .. | .. | 4 | 36 | 57 | 51 | 18 | 1 | .. | .. | 167 |
| | 264–270 | .. | .. | .. | 9 | 37 | 41 | 24 | 10 | 1 | .. | 122 |
| | 271–277 | .. | .. | .. | 1 | 17 | 29 | 22 | 14 | 5 | .. | 88 |
| | 278–284 | .. | .. | .. | 1 | 3 | 4 | 12 | 4 | 3 | 1 | 28 |
| | 285–291 | .. | .. | .. | .. | .. | 1 | 4 | 1 | 5 | .. | 11 |
| | 292–298 | .. | .. | .. | .. | .. | .. | 3 | .. | .. | .. | 3 |
| | *Totals* ........ | 4 | 28 | 73 | 185 | 212 | 167 | 87 | 31 | 14 | 1 | 802 |

6. Correlation table between demand deposits (deposits) and loans, discounts, and overdrafts (loans) for selected national banks in Illinois, for September, 1920. r = 0.638 (*Source: Annual Report of the Comptroller of the Currency*):

Am't $50,000.

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Deposits* | 13 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | 1 | | | | | |
| | 12 | ... | ... | ... | ... | ... | ... | 1 | ... | ... | ... | 1 | ... | ... | 1 | 1 | | |
| | 11 | ... | ... | ... | ... | ... | ... | ... | ... | 1 | 1 | | | | | | | |
| | 10 | ... | ... | ... | ... | ... | ... | ... | ... | ... | 1 | 2 | 1 | | | | | |
| | 9 | ... | ... | ... | ... | ... | ... | ... | ... | 2 | 3 | ... | 1 | ... | 1 | ... | ... | 1 |
| | 8 | ... | ... | ... | ... | ... | 2 | 2 | ... | 1 | 2 | 4 | ... | ... | 1 | | | |
| | 7 | 1 | ... | ... | ... | ... | ... | 2 | 2 | 4 | 2 | 1 | 2 | 2 | | | | |
| | 6 | ... | ... | ... | 1 | ... | 3 | 3 | 3 | 3 | 2 | 1 | 1 | | | | | |
| | 5 | ... | ... | ... | 1 | 1 | 3 | 2 | 7 | 1 | 2 | 3 | 1 | ... | 1 | 1 | | |
| | 4 | ... | ... | ... | 2 | 8 | 5 | 4 | 4 | 4 | 1 | 2 | 2 | ... | ... | 1 | 1 | |
| | 3 | ... | ... | 4 | 4 | 10 | 9 | 5 | 1 | 1 | 1 | ... | 1 | | | | | |
| | 2 | ... | ... | 8 | 8 | 6 | 3 | ... | 1 | 1 | | | | | | | | |
| | 1 | ... | 2 | 2 | 2 | | | | | | | | | | | | | |
| | 0 | ... | ... | 1 | | | | | | | | | | | | | | |
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |

*Loans*

**91. Mean square error of estimate (standard error).**—Whenever the means of the $y$ arrays are approximately on a straight line, let us assume that there exists a linear relation

$$y' = \lambda\frac{\sigma_y}{\sigma_x}x$$

and determine $\lambda$ so that the arithmetic average $(e_y{}^2)$ of the squares of the deviations of the computed value of $y'$ from the actual values of $y$ shall be a minimum. We have

$$\begin{aligned} e_y{}^2 &= \frac{\Sigma(y' - y)^2}{n} \\ &= \frac{1}{n}\sum\left(\lambda^2\frac{\sigma_y{}^2}{\sigma_x{}^2}x^2 - 2\lambda\frac{\sigma_y}{\sigma_x}xy + y^2\right) \\ &= \lambda^2\frac{\sigma_y{}^2}{\sigma_x{}^2}\cdot\frac{1}{n}\sum x^2 - 2\lambda\frac{\sigma_y}{\sigma_x}\frac{\Sigma xy}{n} + \frac{1}{n}\sum y^2 \\ &= \sigma_y{}^2\left[1 - r^2 + (\lambda - r)^2\right], \end{aligned}$$

since $\qquad r = \dfrac{\Sigma xy}{n\sigma_x\sigma_y}$ by definition.

It is clear that $e_y{}^2$ is a minimum if $\lambda = r$. Thus we see that, in the sense of least squares, the best fitting straight line is the line of regression

$$y = r\frac{\sigma_y}{\sigma_x}x$$

If $\lambda = r$, then the sum total of the arithmetic average of the squares of the errors involved is

$$e_y{}^2 = \sigma_y{}^2(1 - r^2)$$

In like manner, we can show that the best fitting straight line, in the sense of least squares to the means of the $x$-arrays, is the line of regression of $x$ on $y$:

$$x = r\frac{\sigma_x}{\sigma_y}y$$

and the mean square of the error involved is

$$e_x{}^2 = \sigma_x{}^2(1 - r^2)$$

For the correlation of price index with monthly tonnage of pig iron, we have

$$e_y = 3.60 \text{ and } e_x = 0.25$$

**92. Error in estimation.**—In §89 we found the equation of the line of regression of $x$ on $y$ to be

$$x = 0.08y + 7.06$$

From this equation we estimated that when the monthly tonnage was 22 millions of tons we could expect an average monthly price index of 8.83. We shall now consider whether this estimate is any better than that resulting from taking the arithmetic average value of the monthly price index, namely

$$\overline{X} = 8.64$$

If the distribution approximates the normal type, the chances are 68 out of 100 that the true value of the variate in question will not differ from the arithmetic average by more than the standard deviation. This means that the chances are 68 out of 100 that the actual price index will be between

$$8.64 - 0.34 = 8.30 \text{ and } 8.64 + 0.34 = 8.98$$

The standard error $e_x$ of $X$ is 0.25. The chances are 68 out of 100 that the error in the estimation from the regression equation will not exceed 0.25. This means that the chances are 68 out of 100 that the actual price index for a monthly production of 22 millions of tons of pig iron will be between

$$8.83 - 0.25 = 8.58 \text{ and } 8.83 + 0.25 = 9.08$$

The error of estimation by using the regression equation has been reduced from 0.34 to 0.25. Thus it is clear that the equation of the line of regression enables us materially to reduce the errors of estimation.

**93. Non-linear regression—Correlation ratio.**—From the expressions for $e_y^2$ and $e_x^2$

$$e_y^2 = \sigma_y^2(1 - r^2),\ e_x^2 = \sigma_x^2(1 - r^2)$$

we find

$$r^2 = 1 - \frac{e_y^2}{\sigma_y^2} = 1 - \frac{e_x^2}{\sigma_x^2}$$

If $e_y$ is small, the dots of the scatter diagram tend to concentrate in a narrow band about the regression line, which may be straight or curved. This suggests the use of $\sqrt{1 - \frac{e_y^2}{\sigma_y^2}}$ as a measure of correlation when the regression is not linear.

This expression is called the *correlation-ratio* of $y$ on $x$, and is written

$$\eta_{yx} = \sqrt{1 - \frac{e_y^2}{\sigma_y^2}}$$

The correlation-ratio of $x$ on $y$ is written

$$\eta_{xy} = \sqrt{1 - \frac{e_x^2}{\sigma_x^2}}$$

When regression is linear,

$$y' = r\frac{\sigma_y}{\sigma_x}x$$

and

$$\eta_{yx} = r$$

When regression is not linear, $e_y^2$ is the mean of the squares of the deviations of $y$ from the line of means of the $y$ arrays. That is, $y' = \bar{y}_x$. This gives us

$$\eta^2{}_{yx} = 1 - \frac{e_y^2}{\sigma_y^2} = \frac{N\sigma_y{}^2 - \Sigma(y - \bar{y}_x)^2}{N\sigma_y{}^2}$$

Let us use the following notation:

$n_x$, frequency within a column;
$S_x$, summation within a column;
$\Sigma$, summation from column to column.

We then have

$$\begin{aligned}
\eta^2{}_{yx} &= \frac{1}{N\sigma_y{}^2}\left\{N\sigma_y{}^2 - \Sigma\left[S_x(y_x - \bar{y}_x)^2\right]\right\}\\
&= \frac{1}{N\sigma_y{}^2}\left\{N\sigma_y{}^2 - \Sigma\left[S_x(y_x{}^2)\right] + 2\Sigma\left[\bar{y}_x \cdot S_x(y_x)\right] - \Sigma\left[S_x(y_x{}^2)\right]\right\}\\
&= \frac{1}{N\sigma_y{}^2}\left\{N\sigma_y{}^2 - \Sigma y^2 + 2\Sigma\left[\bar{y}_x \cdot m_x\bar{y}_x\right] - \Sigma(n_x\bar{y}_x{}^2)\right\}\\
&= \frac{\Sigma(n_x\bar{y}_x{}^2)}{N\sigma_y{}^2}
\end{aligned}$$

If we now change coördinates to the axes $X$ and $Y$, we have

$$\bar{y}_x = \bar{Y} - \bar{Y}_x$$

Whence

$$\eta^2{}_{yx} = \frac{\Sigma[n_x(\bar{Y} - \bar{Y}_x)^2}{N\sigma_y{}^2}$$

In like manner

$$\eta^2{}_{xy} = \frac{\Sigma[n_y(\bar{X} - \bar{X}_y)^2]}{N\sigma_x{}^2}$$

Let us compute the correlation-ratio for the data in table 52, taken from Bulletin No. 280 of the California Agricultural Experiment Station, by S. H. Beckett and R. D. Robertson concerning an experiment on the yield in tons per acre of alfalfa under irrigation.

The reader can verify that $\sigma_y = 2.27$ and $\overline{Y} = 7.48$.

**Table 52.—Yield in Tons per Acre of Alfalfa under Irrigation.**

| *Depth in Inches of Water Applied* | *Yield in Tons per Acre* | | | | | | $\overline{Y}_x$ |
|---|---|---|---|---|---|---|---|
| | *1910* | *1911* | *1912* | *1913* | *1914* | *1915* | *Average* |
| 0 | 3.85 | 5.94 | 5.52 | 2.75 | 2.89 | 2.35 | 3.88 |
| 12 | 4.78 | 7.52 | 6.51 | 4.31 | 5.83 | 4.84 | 5.63 |
| 18 | .... | .... | 7.02 | 5.69 | 8.02 | 6.46 | 6.80 |
| 24 | 6.00 | 8.38 | 8.32 | 6.89 | 9.96 | 7.96 | 7.92 |
| 30 | 7.53 | 9.54 | 9.43 | 7.97 | 11.06 | 8.32 | 8.98 |
| 36 | 7.58 | 9.33 | 9.38 | 8.22 | 12.48 | 8.63 | 9.27 |
| 48 | 8.45 | 9.52 | 8.63 | 8.83 | 10.62 | 8.05 | 9.02 |
| 60 | .... | .... | 10.17 | 7.25 | 10.70 | 5.55 | 8.42 |

**Table 53.—Method of Computation for $N_{yx}$ and $N_{xy}$ for the Data in Table 52**

| $n_x$ | $\overline{Y}_x$ | $\overline{Y} - \overline{Y}_x$ | $(\overline{Y} - \overline{Y}_x)^2$ | $n_x(\overline{Y} - \overline{Y}_x)^2$ |
|---|---|---|---|---|
| 6 | 3.88 | −3.60 | 12.9600 | 77.7600 |
| 6 | 5.63 | −1.85 | 3.4225 | 20.5350 |
| 4 | 6.80 | −0.68 | 0.4624 | 1.8496 |
| 6 | 7.92 | 0.44 | 0.1936 | 1.1616 |
| 6 | 8.98 | 1.50 | 2.2500 | 13.5000 |
| 6 | 9.27 | 1.79 | 3.2041 | 19.2246 |
| 6 | 9.02 | 1.54 | 2.3716 | 14.2296 |
| 4 | 8.42 | 0.94 | 0.8836 | 3.5344 |
| 44 | | | | 151.7948 |

$$\eta^2_{yx} = \frac{151.7948}{44(2.27)^2} = 0.6695;\ \therefore\ \eta_{yx} = 0.82.$$

**Exercises.**

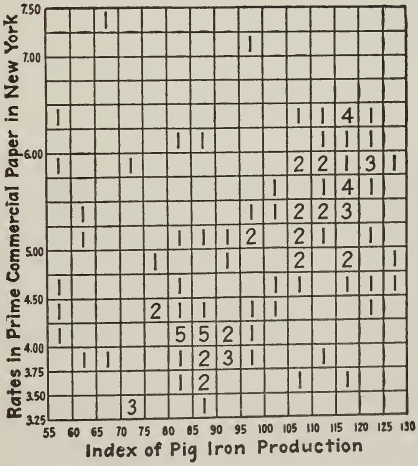
1. Obtain the correlation-ratio for the following data giving the relation between yield of wheat per acre and production cost per bushel on 216

farms in Oklahoma, Kansas, Nebraska, and Missouri in 1920. (*Source: Department Circular 307 of the United States Department of Agriculture*):

**Table 54.—Number of Farms with Yields in Bushels per Acre.**

| Production Cost | *Number of Farms with Yields in Bushels per Acre* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0<br>4.99 | 5<br>9.99 | 10<br>14.99 | 15<br>19.99 | 20<br>24.99 | 25<br>29.99 | 30<br>34.99 | |
| $8–8.99 | 1 | 1 | .. | .. | .. | .. | .. | 2 |
| 7–7.99 | .. | .. | .. | .. | .. | .. | .. | 0 |
| 6–6.99 | 1 | 2 | .. | .. | .. | .. | .. | 3 |
| 5–5.99 | .. | 1 | .. | .. | .. | .. | .. | 1 |
| 4–4.99 | 2 | 2 | 2 | .. | .. | .. | .. | 6 |
| 3–3.99 | .. | 12 | 6 | .. | .. | .. | .. | 18 |
| 2–2.99 | .. | 4 | 42 | 26 | .. | .. | .. | 72 |
| 1–1.99 | .. | 1 | 22 | 49 | 30 | 9 | 2 | 113 |
| 0–0.99 | .. | .. | .. | .. | .. | .. | 1 | 1 |
| Totals........... | 4 | 23 | 72 | 75 | 30 | 9 | 3 | 216 |

2. Obtain the correlation-ratio for the following data on index of pig iron production and rates on prime commercial paper in New York. The numbers in the cells represent the number of months that the index and rate were within the given range. (*Source: The Review of Economic Statistics, Jan., 1925.*)


Rates in Prime Commercial Paper in New York
Index of Pig Iron Production

| | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100 | 105 | 110 | 115 | 120 | 125 | 130 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7.50 | | | 1 | | | | | | | | | | | | | |
| | | | | | | | | | 1 | | | | | | | |
| 7.00 | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | |
| | 1 | | | | | | | | | | 1 | 1 | 4 | 1 | | |
| | | | | | | 1 | 1 | | | | | 1 | 1 | 1 | | |
| 6.00 | 1 | | | 1 | | | | | | | 2 | 2 | 1 | 3 | 1 | |
| | | | | | | | | | | 1 | | 1 | 4 | 1 | | |
| | | 1 | | | | | | | 1 | 1 | 2 | 2 | 3 | | | |
| | | 1 | | | | 1 | 1 | 1 | 2 | | 2 | 1 | | 1 | | |
| 5.00 | | | | | 1 | | | 1 | | | 2 | | 2 | | 1 | |
| | 1 | | | | | 1 | | | | 1 | 1 | | 1 | 1 | 1 | |
| 4.50 | 1 | | | | 2 | 1 | 1 | | 1 | 1 | | | | 1 | | |
| | 1 | | | | | 5 | 5 | 2 | 1 | | | | | | | |
| 4.00 | | 1 | 1 | | | 1 | 2 | 3 | 1 | | | 1 | | | | |
| 3.75 | | | | | | 1 | 2 | | | | 1 | | 1 | | | |
| 3.50 | | | | 3 | | | 1 | | | | | | | | | |
| 3.25 | | | | | | | | | | | | | | | | |

Fig. 43.

**94. Yule's coefficients.**—An observer may note the presence or absence of a specific attribute in each one of a group of individuals. For example, people are sane or insane, blind or not blind, vaccinated or not vaccinated against smallpox; those having smallpox either recover or do not recover. A variable under consideration may not be capable of exact measurement, or may be measured only with difficulty, or we may designedly divide the variates into two or more broad categories. For example, people are tall or short, the temperature is high or low, reaction times are fast or slow. When we note the presence or absence of two attributes, we obtain a fourfold table of the form

| | |
|---|---|
| $a$ | $b$ |
| $c$ | $d$ |

For such a table Yule devised a *coefficient of association* which he[5] denotes by the letter $Q$

$$Q = \frac{ad - bc}{ad + bc}$$

In 1912, Yule[6] devised another coefficient which he called the *coefficient of colligation* ($\omega$)

$$\omega = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Yule expressed his preference for this latter coefficient. The principal merit of either is its simplicity.

Let us compute the coefficient of association for the following data on the influence of previous vaccination in cases of smallpox [*Biometrika,* vol. I, p. 376 ff., W. R. Macdonnell].

[5] G. Udny Yule, "Introduction to the Theory of Statistics," 1917, p. 38. London; Charles Griffin and Company, Limited, Exeter Street, Strand, W. C. 2.

[6] *Journal of the Royal Statistical Soc.*, vol. 75 (1911–12), pp. 579–652.

| | *Recoveries* | *Deaths* | *Totals* |
|---|---|---|---|
| *Vaccinated* | 8,283 | 461 | 8,744 |
| *Unvaccinated* | 1,499 | 822 | 2,321 |
| *Totals* | 9,782 | 1,283 | 11,065 |

$$Q = \frac{8{,}283 \times 822 - 1{,}499 \times 461}{8{,}283 \times 822 + 1{,}499 \times 461} = 0.6561$$

Our conclusion is that there is some relationship between recovery from smallpox and vaccination. This value, 0.6561, of $Q$ is an abstract measure of the degree to which this relationship functioned in this particular instance. It does not mean that approximately 66 per cent of those vaccinated recover. It would seem to indicate that, on the average, 66 per cent of those factors which make for recovery from smallpox are present in those individuals who have been previously vaccinated.

### Exercises.

1. Compute Yule's coefficients for the following data on the influence of previous vaccination in cases of smallpox. (*Source: Biometrika, Vol. I, p. 376ff. W. R. Macdonnell.*)

Sheffield

(a)

| Q = 0.90 | *Recoveries* | *Deaths* | *Totals* |
|---|---|---|---|
| *Vaccinated* | 3,951 | 200 | 4,151 |
| *Unvaccinated* | 278 | 274 | 552 |
| *Totals* | 4,229 | 474 | 4,703 |

(b)

| $Q = 0.88$ | *Mild* | *Severe* | *Totals* |
|---|---|---|---|
| *Vaccinated* | 2,229 | 505 | 2,734 |
| *Unvaccinated* | 229 | 804 | 1,033 |
| *Totals* | 2,458 | 1,309 | 3,767 |

(c) $Q = 0.59$ Glasgow

| *Scars* | *Mild* | *Severe* | *Totals* |
|---|---|---|---|
| *Foveated*...................... | 479 | 24 | 503 |
| *Unfoveated*.................... | 107 | 21 | 128 |
| *Totals*...................... | 586 | 45 | 631 |

(d) $Q = 0.536$

| *Area of Scar* | *Mild* | *Severe* | *Totals* |
|---|---|---|---|
| Over ½ sq. in.................. | 379 | 16 | 395 |
| ½ sq. in. and less.............. | 207 | 29 | 236 |
| *Totals*...................... | 586 | 45 | 631 |

(e) $Q = 0.56$

| *No. of Scars* | *Mild* | *Severe* | *Totals* |
|---|---|---|---|
| 2+........................ | 320 | 16 | 336 |
| 1 .......................... | 266 | 29 | 295 |
| *Totals*...................... | 586 | 45 | 631 |

(f) $Q = 0.614$

| *Years* | *Mild* | *Severe* | *Totals* |
|---|---|---|---|
| 0–20......................... | 244 | 14 | 258 |
| 20+......................... | 1,117 | 268 | 1,385 |
| *Totals*...................... | 1,361 | 282 | 1,643 |

2. Compute Yule's coefficients for the following data obtained from *Biometrika, Vol. XI, p. 23, M. H. Whiting.*

(a)

| | Prison Labor | | |
|---|---|---|---|
| *General Health* | *Hard* | *Light* | *Totals* |
| Good........................ | 260 | 75 | 335 |
| Poor......................... | 33 | 124 | 157 |
| *Totals*...................... | 293 | 199 | 492 |

$Q = 0.857$; $\omega = 0.566$

(b)

| | Muscularity | | |
|---|---|---|---|
| *General Health* | *Muscular* | *Weak* | *Totals* |
| Good | 298 | 37 | 335 |
| Poor | 81 | 77 | 158 |
| *Totals* | 379 | 114 | 493 |

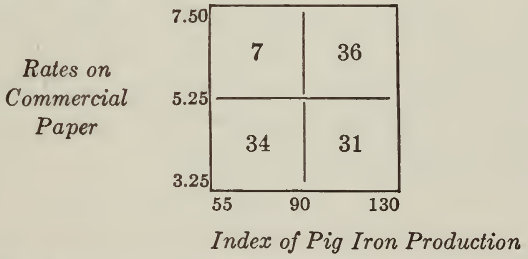
$Q = 0.793;\ \omega = 0.492$

(c)

| | Muscularity | | |
|---|---|---|---|
| *Prison Labor* | *Muscular* | *Weak* | *Totals* |
| Hard | 286 | 7 | 293 |
| Light | 93 | 106 | 199 |
| *Totals* | 379 | 113 | 492 |

$Q = 0.958$

3. Compute Yule's coefficients for the following data on index of pig iron production and rates on prime commercial paper, New York. This table is constructed from data in ex. 2, §93.



**95. Coefficient of concurrent deviations.**—In many problems, we are not interested in the trend or in the deviations from some average. What we *are* interested in is whether a change in one quantity is accompanied by a change in another quantity either in the same or opposite direction. For example, is an increase in price of grain accompanied by an increase in price of eggs? Is an increase in price of wheat accompanied by an increase in price of flour, and that in turn by an increase in price of bread? In computing this coefficient we are not interested in the size of the changes. We need to know only whether they are both in the same direction. The formula for this coefficient is

$$\delta = \pm \sqrt{\pm \frac{2c - n}{n}}$$

where $n$ is one less than the number of pairs of items or the total number of deviations; $c$ is the number of concurrent deviations.

The application of the formula can be shown from the data in table 55. For $c$ we shall consider changes in opposite direction. Thus, the production in 1892 decreased while the price increased. This counts for one value of $c$. The production in 1893 increased over 1892, while the price declined. This counts for the second value of $c$. We find for the period given that $c = 8$. Then

**Table 55.—Yearly Price and Production of Cotton.**

| *Year* | *Cotton Production in Millions of Lbs.* | *Price per Lb. in Cents* | *Year* | *Cotton Production in Millions of Lbs.* | *Price per Lb. in Cents* |
|---|---|---|---|---|---|
| 1891 | 4,450 | 7.3 | 1896 | 4,250 | 7.3 |
| 1892 | 3,350 | 8.4 | 1897 | 5,500 | 5.6 |
| 1893 | 3,700 | 7.5 | 1898 | 5,700 | 4.9 |
| 1894 | 5,000 | 5.9 | 1899 | 4,750 | 7.6 |
| 1895 | 3,550 | 8.2 | 1900 | 5,150 | 9.3 |
| | | | 1901 | 4,850 | 8.1 |

$$\delta = -\sqrt{\frac{16 - 10}{10}} = -0.77$$

We use $n = 10$, for the data given do not enable us to measure the deviation for 1891. The same result would be reached by counting the number $c = 2$ of deviations in the same direction. Then $2c - 10 = 4 - 10 = -6$. We would then make use of the negative sign under the radical and also the negative sign before the radical.

Thus,

$$\delta = -\sqrt{-\frac{4 - 10}{10}} = -0.77$$

### Exercises.

Compute the coefficient of concurrent deviations for the following data on the production and price of cotton and the production and price of corn (*Source: Yearbook of the Department of Agriculture, 1922, 1925*):

**Table 56.—Yearly Average Farm Price and Production of Cotton and Corn.**

| *Year* | COTTON<br>*1,000 Bales of 500 Lbs. Gross Wt.* | *Average Farm Price per Lb. Dec. 1* | CORN<br>*Production in 1,000 Bu.* | *Average Farm Price per Bu. Dec. 1* |
|---|---|---|---|---|
| 1901 | 9,510 | 20.0 | 1,613,528 | 60.1 |
| | 10,631 | 7.6 | 2,619,499 | 40.1 |
| | 9,851 | 10.5 | 2,346,897 | 42.1 |
| | 13,438 | 9.0 | 2,528,662 | 43.7 |
| | 10,575 | 10.8 | 2,748,949 | 40.8 |
| 1906 | 13,274 | 9.6 | 2,897,662 | 39.3 |
| | 11,107 | 10.4 | 2,512,065 | 50.9 |
| | 13,242 | 8.7 | 2,544,957 | 60.0 |
| | 10,005 | 13.9 | 2,572,336 | 58.6 |
| | 11,609 | 14.1 | 2,886,260 | 48.0 |
| 1911 | 15,693 | 8.8 | 2,531,488 | 61.8 |
| | 13,703 | 11.9 | 3,124,746 | 48.7 |
| | 14,156 | 12.2 | 2,446,988 | 69.1 |
| | 16,135 | 6.8 | 2,672,804 | 64.4 |
| | 11,192 | 11.3 | 2,994,793 | 57.5 |
| 1916 | 11,450 | 19.6 | 2,566,927 | 88.9 |
| | 11,302 | 27.7 | 3,065,233 | 127.9 |
| | 12,041 | 27.6 | 2,502,665 | 136.5 |
| | 11,421 | 35.6 | 2,811,302 | 134.5 |
| | 13,440 | 13.9 | 3,208,584 | 67.0 |
| 1921 | 7,954 | 16.2 | 3,068,569 | 42.3 |
| | 9,762 | 23.8 | 2,906,020 | 65.8 |
| | 10,140 | 31.0 | 3,053,557 | 72.6 |
| | 13,628 | 22.6 | 2,342,745 | 98.2 |
| | 16,106 | 18.2 | 2,900,581 | 67.4 |

*Ans.* $\delta = -0.76$, the same for both.

**96. Correlation from ranks.**—Let us suppose a group of individuals ranked with respect to two different abilities, say ability in mathematics and ability in Latin. It is clear that such a record constitutes an ordinary correlation table wherein, however, a definite measure is not given to each variate. Equal difference in rank does not constitute equal difference in measure of an ability. A degree of correlation can be computed by using Pearson's sum-product method. We are going to show in this section how to compute this value of $r$ by a shorter procedure. We first need some preliminary formulas.

Let $N$ be the number of ranks, the same for both variables. Let $\bar{x}$ and $\bar{y}$ be the mean rank of the respective variables. Then

$$\bar{x} = \bar{y} = \frac{N+1}{2}$$

Let $x$ and $y$ be the ranks of the same individual with respect to the two abilities. Then[7]

$$\begin{aligned} N\sigma_x^2 &= \sum_1^N (x - \bar{x})^2 \\ &= \Sigma x^2 - 2\bar{x}\Sigma x + \Sigma \bar{x}^2 \\ &= \frac{N(N+1)(2N+1)}{6} - 2\frac{N+1}{2}\cdot\frac{N(N+1)}{2} + N\left(\frac{N+1}{2}\right)^2 \\ &= \tfrac{1}{12}(N^3 - N) = \tfrac{1}{12}N(N^2 - 1) \end{aligned}$$

Now

$$(x - y)^2 = x^2 + y^2 - 2xy$$

and

$$\Sigma(x - y)^2 = N\sigma^2_{x-y},$$

by definition of standard deviation.

Therefore
$$\begin{aligned} N\sigma^2_{x-y} &= \Sigma x^2 + \Sigma y^2 - 2\Sigma xy \\ &= N\sigma_x^2 + N\sigma_y^2 - 2rN\sigma_x\sigma_y \end{aligned}$$

since
$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$$

Therefore
$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma^2_{x-y}}{2\sigma_x\sigma_y}$$

[7] $\Sigma x = 1 + 2 + \cdots + N = N\frac{N+1}{2}$; $\Sigma x^2 = 1 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}$

See Rietz-Crathorne, "College Algebra," p. 92, revised edition, Henry Holt & Co.

But $\sigma_x = \sigma_y$, for we have the same number of ranks in each case, and they have the same mean.

Therefore
$$r = 1 - \frac{\sigma^2{}_{x-y}}{2\sigma_x{}^2}$$
$$= 1 - \frac{\Sigma(x-y)^2}{2N\sigma_x{}^2}$$
$$= 1 - \frac{\Sigma(x-y)^2}{2 \cdot \frac{1}{12}(N^3 - N)}$$
$$r = 1 - \frac{6\Sigma(x-y)^2}{N(N^2-1)}$$

The correlation from ranks computed from this formula must agree with Pearson's sum-product correlation coefficient $r = \frac{\Sigma xy}{N\sigma_x\sigma_y}$, where $x$ and $y$ are the ranks of the individuals. The value of $r$ computed from the sum-product formula, where for $x$ and $y$ we use the measurements of the variates instead of their ranks, in general is not equal to the correlation from ranks. For this reason, it is customary to denote by $\rho$ (rho) the correlation from ranks. Then

$$\rho = 1 - \frac{6\Sigma(x-y)^2}{N(N^2-1)} \tag{4}$$

The following example shows that the actual correlation between two series can be made to change very much without changing ranks. Thus, the two hypothetical series of grades in mathematics and chemistry:

| | | | | | |
|---|---|---|---|---|---|
| *Mathematics:* | 100 | 99 | 98 | 97 | 6 |
| *Chemistry:* | 100 | 95 | 70 | 61 | 60 |

give a perfect correspondence in rank, but the correlation is far from perfect.

The form for the computation of $\rho$ is illustrated by the following data (table 57), which give the corresponding grades of 60 different students at Colorado College in Trigonometry and Latin. Wherever ties in ranks occur we have used the bracket rank method explained in article 97.

Table 57.—Form for Computation of $\rho$. Corresponding Grades of 60 Students at Colorado College in Trigonometry and Latin.

| *Trig.* | $x$ | *Latin* | $y$ | $x - y$ | $(x - y)^2$ |
|---|---|---|---|---|---|
| 93 | 1 | 82 | 18 | −17 | 289 |
| 90 | 2 | 90 | 6 | −4 | 16 |
| 90 | 2 | 80 | 21 | −19 | 361 |
| 88 | 4 | 85 | 14 | −10 | 100 |
| 88 | 4 | 91 | 4 | 0 | 0 |
| 86 | 6 | 94 | 2 | 4 | 16 |
| 85 | 7 | 80 | 21 | −14 | 196 |
| 85 | 7 | 67 | 42 | −35 | 1,225 |
| 85 | 7 | 90 | 6 | 1 | 1 |
| 85 | 7 | 64 | 51 | −44 | 1,936 |
| 85 | 7 | 88 | 12 | −5 | 25 |
| 85 | 7 | 80 | 21 | −14 | 196 |
| 84 | 13 | 88 | 12 | 1 | 1 |
| 84 | 13 | 78 | 27 | −14 | 196 |
| 84 | 13 | 78 | 27 | −14 | 196 |
| 83 | 16 | 65 | 45 | −29 | 841 |
| 83 | 16 | 90 | 6 | 10 | 100 |
| 82 | 18 | 80 | 21 | −3 | 9 |
| 82 | 18 | 90 | 6 | 12 | 144 |
| 80 | 20 | 90 | 6 | 14 | 196 |
| 78 | 21 | 60 | 54 | −33 | 1,089 |
| 78 | 21 | 73 | 32 | −11 | 121 |
| 78 | 21 | 65 | 45 | −24 | 576 |
| 77 | 24 | 96 | 1 | 23 | 529 |
| 77 | 24 | 91 | 4 | 20 | 400 |
| 75 | 26 | 90 | 6 | 20 | 400 |
| 75 | 26 | 92 | 3 | 23 | 529 |
| 75 | 26 | 70 | 35 | −9 | 81 |
| 75 | 26 | 70 | 35 | −9 | 81 |
| 75 | 26 | 70 | 35 | −9 | 81 |
| 73 | 31 | 70 | 35 | −4 | 16 |
| 72 | 32 | 42 | 60 | −28 | 784 |
| 71 | 33 | 83 | 16 | 17 | 289 |
| 68 | 34 | 70 | 35 | −1 | 1 |
| 68 | 34 | 65 | 45 | −11 | 121 |
| 68 | 34 | 68 | 41 | −7 | 49 |
| 66 | 37 | 73 | 32 | 5 | 25 |
| 66 | 37 | 65 | 45 | −8 | 64 |
| 66 | 37 | 65 | 45 | −8 | 64 |
| 65 | 40 | 65 | 45 | 5 | 25 |
| 65 | 40 | 70 | 35 | 5 | 25 |
| 65 | 40 | 60 | 54 | −14 | 196 |
| 63 | 43 | 81 | 20 | 23 | 529 |
| 63 | 43 | 78 | 27 | 16 | 256 |
| 62 | 45 | 84 | 15 | 30 | 900 |
| 60 | 46 | 83 | 16 | 30 | 900 |
| 60 | 46 | 80 | 21 | 25 | 625 |
| 60 | 46 | 66 | 43 | 3 | 9 |
| 60 | 46 | 80 | 21 | 25 | 625 |
| 60 | 46 | 82 | 18 | 28 | 784 |
| 60 | 46 | 60 | 54 | −8 | 64 |
| 57 | 52 | 62 | 53 | −1 | 1 |
| 55 | 53 | 74 | 31 | 22 | 484 |
| 55 | 53 | 72 | 34 | 19 | 361 |
| 55 | 53 | 60 | 54 | −1 | 1 |
| 55 | 53 | 57 | 59 | −6 | 36 |
| 48 | 57 | 78 | 27 | 30 | 900 |
| 45 | 58 | 64 | 51 | 7 | 49 |
| 40 | 59 | 66 | 43 | 16 | 256 |
| 40 | 59 | 60 | 54 | 5 | 25 |

$$\Sigma(x - y)^2 = 18{,}395$$

$$\rho = 1 - \frac{6 \times 18{,}395}{60 \times 3{,}599} = 0.49$$

**97. Ties in ranks.**—The only uncertainty in the application of this formula comes when two or more variates have the same measurement. In this case the rank to be assigned to the variates is determined by one of the two following plans:

1. *The bracket rank method.*—The variates having the same measurement are assigned the same rank, which is the number next greater than the rank of the variate immediately preceding the ties. The next variate after the ties is given the same rank that it would have had in case there had been no tie.

2. *The mid-rank method.*—The variates having the same measurement are given the same rank. The next variate after the ties is given the same rank it would have had in case there had been no tie. The rank of those variates which are tied is a number midway between the rank of that variate which immediately precedes and the rank of that variate which immediately follows the tied variates. Under this method the sum of the ranks is the same as if there had been no ties. Table 58 illustrates both methods.

**Table 58.—Method of Assigning Ranks in Case of Ties in Rank.**

| *State* | *% Illiteracy* | *Bracket Method* | *Mid-rank Method* |
|---|---|---|---|
| Iowa | 1.1 | 1 | 1 |
| Nebr. | 1.4 | 2 | 2 |
| Idaho | 1.5 | 3 | 3.5 |
| Oregon | 1.5 | 3 | 3.5 |
| Minn. | 1.8 | 5 | 5 |
| Vermont | 3.0 | 6 | 7 |
| Mich. | 3.0 | 6 | 7 |
| Missouri | 3.0 | 6 | 7 |
| Mass. | 4.6 | 9 | 9 |

### Exercises.

1. Table 59 gives the rank of the 49 states of the United States with respect to (A) the per cent of illiteracy in the population 10 years of age and over, as given by the 14th census, 1920, vol. II, p. 1154; (B) the proportion of children of both sexes 10 to 15 years of age engaged in gainful occupations as compiled from data given in the 14th census, 1920, vol. IV, p. 514. Compute $\rho$.

**Table 59.—Rank of the 49 States of the United States with Respect to (A) Per Cent of Illiteracy in Population 10 Years of Age and Over, and (B) Proportion of Children of Both Sexes Engaged in Gainful Occupations.**

| | *A* | % | *B* | % | | *A* | % | *B* | % |
|---|---|---|---|---|---|---|---|---|---|
| Iowa | 1 | 1.1 | 15 | 3.37 | Penn | 26 | 4.6 | 28 | 5.6 |
| Neb | 2 | 1.4 | 16 | 3.38 | Mass | 27 | 4.7 | 38 | 8.6 |
| Idaho | 3 | 1.5 | 3 | 2.94 | N. Y | 28 | 5.1 | 23 | 4.7 |
| Oregon | 4 | 1.5+ | 7 | 3.02 | N. J | 29 | 5.1+ | 33 | 7.6 |
| Kansas | 5 | 1.6 | 18 | 3.43 | Maryland | 30 | 5.6 | 32 | 7.5 |
| So. Dak | 6 | 1.7 | 11 | 3.26 | Nevada | 31 | 5.9 | 2 | 2.4 |
| Wash | 7 | 1.7+ | 14 | 3.35 | Del | 32 | 5.9+ | 30 | 5.8 |
| Minn | 8 | 1.8 | 6 | 2.984 | Conn | 33 | 6.2 | 35 | 8.07 |
| Utah | 9 | 1.9 | 20 | 3.9 | W. V | 34 | 6.4 | 19 | 3.8 |
| Wyoming | 10 | 2.1 | 5 | 2.980 | R. I | 35 | 6.5 | 43 | 13.4 |
| N. Dak | 11 | 2.1+ | 10 | 3.20 | Tex | 36 | 8.3 | 42 | 12.58 |
| Indiana | 12 | 2.2 | 25 | 5.22 | Kentucky | 37 | 8.4 | 37 | 8.3 |
| Montana | 13 | 2.3 | 1 | 2.3 | Ark | 38 | 9.4 | 45 | 18.5 |
| Wisconsin | 14 | 2.4 | 24 | 5.1 | Fla | 39 | 9.6 | 39 | 8.7 |
| Ohio | 15 | 2.8 | 8 | 3.04 | Tenn | 40 | 10.3 | 40 | 12.2 |
| Dist. of Columbia | 16 | 2.8+ | 27 | 5.31 | Va | 41 | 11.2 | 36 | 8.17 |
| Vt | 17 | 3.0− | 12 | 3.31 | N. C | 42 | 13.1 | 44 | 16.6 |
| Mich | 18 | 3.0 | 17 | 3.42 | Ariz | 43 | 15.3 | 31 | 7.1 |
| Missouri | 19 | 3.0+ | 29 | 5.7 | Ga | 44 | 15.3+ | 46 | 20.7 |
| Colo | 20 | 3.2 | 21 | 4.3 | N. M | 45 | 15.6 | 22 | 4.5 |
| Cal | 21 | 3.3 | 4 | 2.97 | Ala | 46 | 16.1 | 47 | 24.1 |
| Maine | 22 | 3.3+ | 9 | 3.12 | Miss | 47 | 17.2 | 49 | 25.4 |
| Ill | 23 | 3.4 | 26 | 5.28 | S. C | 48 | 18.1 | 48 | 24.4 |
| Okla | 24 | 3.8 | 34 | 7.8 | La | 49 | 21.9 | 41 | 12.51 |
| N. Hamp | 25 | 4.4 | 13 | 3.34 | | | | | |

*Fourteenth Census of U. S., 1920, Vol. II, pp. 514, 1154.*

$$\rho = 1 - \frac{6 \times 4{,}384}{117{,}600} = 0.776$$

2. Table 60 gives the rank of each of the states with respect to (A) prisoners and juvenile delinquents enumerated on Jan. 1, 1910, and (B) percentage of population 10 to 15 years of age engaged in gainful occupation.

**Table 60.—Rank of Each of the States with Respect to (A) Prisoners and Juvenile Delinquents Enumerated on Jan. 1, 1910, and (B) Percentage of Population 10 to 15 Years of Age Engaged in Gainful Occupations.**

| | *A* | *Actual No. per* 100,000 | *B* | % | | *A* | *Actual Number per* 100,000 | *B* | % |
|---|---|---|---|---|---|---|---|---|---|
| N. J. .......... | 1 | 201 | 24 | 9.5 | Idaho .......... | 26 | 410 | 14 | 7.14 |
| Ill. ............ | 2 | 202 | 21 | 9.0 | Colo. .......... | 27 | 415 | 13 | 7.06 |
| R. I. ........... | 3 | 207 | 33 | 14.4 | N. C. .......... | 28 | 445 | 46 | 45.8 |
| Conn. .......... | 4 | 222 | 26 | 10 | Tex. ........... | 29 | 470 | 43 | 32.1 |
| Penn. .......... | 5 | 222.7 | 30 | 11.5 | S. D. .......... | 30 | 498 | 29 | 11.3 |
| N. Y. .......... | 6 | 223.8 | 11 | 6.9 | Md. ............ | 31 | 505 | 35 | 15.6 |
| Ohio ........... | 7 | 224.3 | 18 | 8.4 | La. ............ | 32 | 506 | 40 | 26.1 |
| Wis. ........... | 8 | 228 | 17 | 8.1 | Okla. .......... | 33 | 521 | 37 | 20.8 |
| Minn. .......... | 9 | 236 | 15 | 7.2 | Tenn. .......... | 34 | 615 | 42 | 28.8 |
| Ore. ........... | 10 | 240 | 5 | 5.2 | Ky. ............ | 35 | 641 | 38 | 21.3 |
| Neb. ........... | 11 | 254 | 19 | 8.5 | N. D. .......... | 36 | 665 | 27 | 10.4 |
| Mass. .......... | 12 | 262 | 25 | 9.7 | Wy. ............ | 37 | 667 | 8 | 6.7 |
| N. H. .......... | 13 | 277 | 20 | 8.7 | Ark. ........... | 38 | 672 | 44 | 43.2 |
| Iowa ........... | 14 | 279 | 22 | 9.1 | S. C. .......... | 39 | 751 | 48 | 51.8 |
| Maine .......... | 15 | 281 | 12 | 7.02 | Va. ............ | 40 | 759 | 39 | 21.9 |
| Mich. .......... | 16 | 282 | 7 | 6.2 | Mont. .......... | 41 | 789 | 4 | 5.0 |
| Wash. .......... | 17 | 311 | 3 | 4.9 | W. V. .......... | 42 | 796 | 34 | 15.1 |
| Mo. ............ | 18 | 318 | 31 | 13.5 | Fla. ........... | 43 | 883 | 41 | 26.3 |
| Cal. ........... | 19 | 333 | 6 | 5.4 | N. M. .......... | 44 | 914 | 32 | 14 |
| Vt. ............ | 20 | 336 | 9 | 6.82 | Ga. ............ | 45 | 942 | 45 | 43.3 |
| Ind. ........... | 21 | 347 | 28 | 11 | Ala. ........... | 46 | 1,042 | 47 | 51.6 |
| D. C. .......... | 22 | 358 | 1 | 4.6 | Ariz. .......... | 47 | 1,098 | 16 | 7.6 |
| Utah ........... | 23 | 372 | 10 | 6.83 | Miss. .......... | 48 | 1,102 | 49 | 53.3 |
| Kans. .......... | 24 | 399 | 23 | 9.2 | Nev. ........... | 49 | 2,156 | 2 | 4.8 |
| Del. ........... | 25 | 402 | 36 | 15.8 | | | | | |

*Department of Commerce, Bureau of Census, Bull. 121, p. 114, and Thirteenth Census of U. S., 1910, Vol. IV, p. 75.*

Compute $\rho$ for: *Ans.*

(a) All of the states. 0.39

(b) All states except Wy., Mont., Ariz., Nev. 0.65

(c) The states east of the Mississippi River. 0.72

(d) The states east of Colorado. 0.72

(e) All states except Nev., Idaho, Utah, Ariz., N. M., Wy., Mont., Colo. 0.69

**98. Spearman's foot rule.**—There has been devised an empirical formula called Spearman's foot rule for the computation of correspondence from ranks:

$$R = 1 - \frac{6\Sigma g}{N^2 - 1}$$

In this formula $g$ denotes a positive difference in rank. This formula is easy to compute but gives only a rough estimate of the measure of correlation.

Spearman's foot rule ordinarily is not used if more than thirty cases are involved. This rule is often used in educational work where groups of thirty or fewer students are involved, and the rank of the students is known with respect to two abilities.

### Exercises.

1. Compute $R$ for the data in table 61 for the combinations given below.

   (a) Mathematics and English.
   (b) Thorndike intelligence test and mathematics.
   (c) Thorndike intelligence test and English.

## Multiple and Partial Correlation.

**99. Simple correlation.**—In table 61 is given the grade, reduced to percentages, made by each of thirty freshmen at Colorado College in Sept., 1925, on the Thorndike Intelligence test. The same table gives the grade made by the same freshmen in their college mathematics and English. By methods already described, the simple correlation coefficient can be computed between any two of the sets of data. The two lines of regression for any two sets can also be computed. The results for the combinations indicated are as follows:

| | | |
|---|---|---|
| $(X_1X_2)$ | $X_1 = 66.389 + 0.1237X_2$ | |
| | $X_2 = 19.08 + 0.739X_1$ | $r_{12} = 0.302$ |
| $(X_1X_3)$ | $X_1 = 57.141 + 0.2232X_3$ | |
| | $X_3 = 60.5 + 0.297X_1$ | $r_{13} = 0.257$ |
| $(X_2X_3)$ | $X_2 = 51.835 + 0.2791X_3$ | |
| | $X_3 = 36.5 + 0.6203X_2$ | $r_{23} = 0.416$ |

In the computation of these results (for example, the simple correlation coefficient between $X_1$ and $X_2$) the presence and influence of the remaining variables (in this case $X_3$) are ignored. It is as though a chemist, in studying some reaction between two chemicals, ignored the presence and influence of a third chemical in his test tubes.

**Table 61.—Corresponding Marks in Mathematics, English, and Thorndike Intelligence Test.**

| *Thorndike Intelligence* | *Math.* | *English* | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | $X_2$ | $X_3$ | $X_1^2$ | $X_2^2$ | $X_3^2$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ |
| 81 | 70 | 88 | 6,561 | 4,900 | 7,744 | 5,670 | 7,128 | 6,160 |
| 74 | 85 | 87 | 5,476 | 7,225 | 7,569 | 6,290 | 6,438 | 7,395 |
| 68 | 60 | 75 | 4,624 | 3,600 | 5,625 | 4,080 | 5,100 | 4,500 |
| 78 | 75 | 81 | 6,084 | 5,625 | 6,561 | 5,850 | 6,318 | 6,075 |
| 71 | 75 | 85 | 5,041 | 5,625 | 7,225 | 5,325 | 6,035 | 6,375 |
| 77 | 77 | 77 | 5,929 | 5,929 | 5,929 | 5,929 | 5,929 | 5,929 |
| 75 | 81 | 88 | 5,625 | 7,744 | 7,744 | 6,075 | 6,600 | 7,128 |
| 71 | 75 | 75 | 5,041 | 5,625 | 5,625 | 5,325 | 5,325 | 5,625 |
| 73 | 40 | 81 | 5,329 | 1,600 | 6,561 | 2,920 | 5,913 | 3,240 |
| 70 | 91 | 81 | 4,900 | 8,281 | 6,561 | 6,370 | 5,670 | 7,371 |
| 67 | 72 | 75 | 4,489 | 5,184 | 5,625 | 4,824 | 5,025 | 5,400 |
| 78 | 85 | 91 | 6,084 | 7,225 | 8,281 | 6,630 | 7,098 | 7,735 |
| 79 | 91 | 87 | 6,241 | 8,281 | 7,569 | 7,189 | 6,873 | 7,917 |
| 68 | 65 | 75 | 4,624 | 4,225 | 5,625 | 4,420 | 5,100 | 4,875 |
| 77 | 80 | 80 | 5,929 | 6,400 | 6,400 | 6,160 | 6,160 | 6,400 |
| 66 | 77 | 91 | 4,356 | 5,929 | 8,281 | 5,082 | 6,006 | 7,007 |
| 82 | 91 | 87 | 6,724 | 8,281 | 7,569 | 7,462 | 7,134 | 7,917 |
| 77 | 81 | 60 | 5,929 | 6,561 | 3,600 | 6,237 | 4,620 | 4,860 |
| 73 | 90 | 91 | 5,329 | 8,100 | 8,281 | 6,570 | 6,643 | 8,190 |
| 75 | 77 | 75 | 5,625 | 5,929 | 5,625 | 5,775 | 5,625 | 5,775 |
| 68 | 40 | 85 | 4,624 | 1,600 | 7,225 | 2,720 | 5,780 | 3,400 |
| 68 | 70 | 85 | 4,624 | 4,900 | 7,225 | 4,760 | 5,780 | 5,950 |
| 85 | 90 | 85 | 7,225 | 8,100 | 7,225 | 7,650 | 7,225 | 7,650 |
| 83 | 85 | 90 | 6,889 | 7,225 | 8,100 | 7,055 | 7,470 | 7,650 |
| 83 | 70 | 90 | 6,889 | 4,900 | 8,100 | 5,810 | 7,470 | 6,300 |
| 84 | 73 | 86 | 7,056 | 5,329 | 7,396 | 6,132 | 7,224 | 6,278 |
| 84 | 84 | 83 | 7,056 | 7,056 | 6,889 | 7,056 | 6,972 | 6,972 |
| 82 | 60 | 87 | 6,724 | 3,600 | 7,569 | 4,920 | 7,134 | 5,220 |
| 72 | 70 | 86 | 5,184 | 4,900 | 7,396 | 5,040 | 6,192 | 6,020 |
| 81 | 70 | 83 | 6,561 | 4,900 | 6,889 | 5,670 | 6,723 | 5,810 |
| 2,270 | 2,250 | 2,490 | 172,772 | 174,779 | 208,014 | 170,996 | 188,710 | 187,124 |

**100. Multiple regression.**—Assume a linear relationship among the variables. We have

$$X_1 = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \tag{5}$$

Substituting in this equation the values given in table 61, we have

$$\begin{aligned} 81 &= a + b_{12\cdot3}70 + b_{13\cdot2}88 \\ 74 &= a + b_{12\cdot3}85 + b_{13\cdot2}87 \\ 68 &= a + b_{12\cdot3}60 + b_{13\cdot2}75 \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned} \tag{6}$$

$$2{,}270 = 30a + b_{12\cdot3}2{,}250 + b_{13\cdot2}2{,}490 \tag{7}$$

We obtain thus, in all, thirty equations. Adding, we have equation (7), which in the general case may be written

$$\Sigma X_1 = Na + b_{12\cdot3}\Sigma X_2 + b_{13\cdot2}\Sigma X_3 \tag{8}$$

Multiply each equation in set (6) by the coefficient of $b_{12\cdot3}$ in that equation and add the thirty new equations. We have

$$\begin{aligned} 5{,}670 &= 70a + b_{12\cdot3}(70)^2 + b_{13\cdot2}(88)(70) \\ (74)(85) &= 85a + b_{12\cdot3}(85)^2 + b_{13\cdot2}(87)(85) \\ (68)(60) &= 60a + b_{12\cdot3}(60)^2 + b_{13\cdot2}(75)(60) \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned} \tag{9}$$

$$170{,}996 = 2{,}250a + 174{,}779b_{12\cdot3} + 187{,}124b_{13\cdot2} \tag{10}$$

We obtain thus thirty equations in set (9). Adding, we have equation (10), which in the general case may be written

$$\Sigma(X_1X_2) = a\Sigma X_2 + b_{12\cdot3}\Sigma X_2^2 + b_{13\cdot2}\Sigma(X_2X_3) \tag{11}$$

Multiply each equation in set (6) by the coefficient of $b_{13\cdot2}$ in that equation and add the thirty new equations. We have

$$\begin{aligned} (81)(88) &= 88a + 6{,}160b_{12\cdot3} + 7{,}744b_{13\cdot2} \\ (74)(87) &= 87a + (85)(87)b_{12\cdot3} + (87)^2b_{13\cdot2} \\ (68)(75) &= 75a + (60)(75)b_{12\cdot3} + (75)^2b_{13\cdot2} \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned} \tag{12}$$

$$188{,}710 = 2{,}490a + 187{,}124b_{12\cdot3} + 208{,}014b_{13\cdot2} \tag{13}$$

There are thirty equations in set (12). Adding, we have equation (13), which in the general case may be written

$$\Sigma(X_1X_3) = a\Sigma X_3 + b_{12\cdot3}\Sigma X_2X_3 + b_{13\cdot2}\Sigma X_3^2 \tag{14}$$

The method of procedure for more variables should be obvious.

The solution of equations (7), (10), (13) for $a$, $b_{12\cdot3}$, $b_{13\cdot2}$ gives

$$a = 51.771;\quad b_{12\cdot3} = 0.111;\quad b_{13\cdot2} = 0.188$$

Hence

(15) $$X_1 = 51.771 + 0.111X_2 + 0.188X_3$$

Proceeding in like manner, we find

(16) $$X_2 = 11.687 + 0.704X_1 + 0.121X_3$$

(17) $$X_3 = 59.94 + 0.277X_1 + 0.028X_2$$

**101. Estimates.**—Equations (15), (16), and (17) are used to estimate the average value of $X_1$, $X_2$, and $X_3$, when values of the other variables are known.

Thus (15) shows that the average grade on the Thorndike test for those who obtain a zero grade in both mathematics and English is 50.178, whereas the average grade on the Thorndike test for those who obtain 100 in both mathematics and English is 50.178 + 11.1 + 18.8 = 80.1.

Equation (16) shows that the average grade in mathematics for those who score zero in both the Thorndike test and English is 11.687. If $X_1 = 100$ and $X_3 = 0$, then $X_2 = 82.08$. If $X_1 = 0$ and $X_3 = 100$, then $X_2 = 23.787$.

Equation (17) shows that the average grade in English for those who score zero in both the Thorndike test and in mathematics is 59.94. If $X_1 = 100$ and $X_2 = 0$, then $X_3 = 87.6$. If $X_1 = 0$ and $X_2 = 0$, then $X_3 = 62.7$.

**102a. Multiple correlation.**—Let $d$ represent the deviation of the computed from the actual value; then

$$d = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 - X_1$$

Multiply both sides of this equation by $d$ and add. We find

$$\Sigma d^2 = a\Sigma d + b_{12\cdot3}\Sigma dX_2 + b_{13\cdot2}\Sigma dX_3 + \Sigma dX$$

But[8]

$$\Sigma d = \Sigma dX_2 = \Sigma dX_3 = 0$$

[8] The easiest way to prove this is to use the calculus as follows:

$$f \equiv \Sigma d^2 = \Sigma(a + b_{12.3}X_2 + b_{13\cdot2}X_3 - X_1)^2$$

In order that $f$ may be a minimum we must have

$$\frac{\partial f}{\partial a} = 2\Sigma(a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 - X_1) = 2\Sigma d = 0$$

$$\therefore\ \Sigma d = 0$$

$$\frac{\partial f}{\partial b_{12\cdot3}} = 2\Sigma(a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 - X_1)X_2 = 2\Sigma dX_2 = 0$$

$$\therefore \Sigma dX_2 = 0$$

$$\frac{\partial f}{\partial b_{13\cdot2}} = 2\Sigma(a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 - X_1)X_3 = 2\Sigma dX_3 = 0$$

$$\therefore\ \Sigma dX_3 = 0$$

Then

$$\begin{aligned}\Sigma d^2 &= -\Sigma dX_1\\ &= -\Sigma X_1^2 - a\Sigma X_1 - b_{12\cdot 3}\Sigma X_1X_2 - b_{13\cdot 2}\Sigma X_1X_3\end{aligned}$$

If we let $e^2{}_{1\cdot 23}$ represent the average value of the squares of these deviations, then

$$e^2{}_{1\cdot 23} = \frac{\Sigma d^2}{N} = \frac{-\Sigma X_1^2 - a\Sigma X_1 - b_{12\cdot 3}\Sigma X_1X_2 - b_{13\cdot 2}\Sigma X_1X_3}{N}$$

Let us call $e_{1\cdot 23}$ the standard error of $X_1$. If $e_{1\cdot 23}$ is as large as $\sigma_1$ (the standard deviation of $X_1$), then equation (15) does not give any better estimate of $X_1$ than does $\overline{X}_1$. If, however, $e_{1\cdot 23}$ is smaller than $\sigma_1$, the equation gives a better estimate of $X_1$ than $\overline{X}_1$ does. In computing $\overline{X}_1$ and $\sigma_1$ only the recorded values of $X_1$ are used. In computing $X_1$ from the equation and in computing $e_{1\cdot 23}$, recorded values of $X_1$, $X_2$, and $X_3$ are used. If the equation gives a better value for $X_1$, it must be due to the influence of $X_2$ and $X_3$. The significance of this influence is indicated by the relation between the standard error and the standard deviation. Both are in absolute terms. An abstract measure of the relationship is obtained by dividing the standard error by the standard deviation.

We use as a measure of the multiple correlation $R_{1\cdot 23}$ between $X_1$ and the variables $X_2$, $X_3$ the following:

$$R_{1\cdot 23} = \sqrt{1 - \frac{e^2{}_{1\cdot 23}}{\sigma_1{}^2}}$$

Observe the analogy between this formula for the coefficient of multiple correlation and the formulas for simple correlation·

$$r = \sqrt{1 - \frac{e_x{}^2}{\sigma_x{}^2}} = \sqrt{1 - \frac{e_y{}^2}{\sigma_y{}^2}}$$

For the illustrative problem we have

$$e^2{}_{1\cdot 23} = \frac{172{,}772 - (51.771)(2{,}270) - (0.111)(170{,}996) - (0.188)(188{,}710)}{30}$$

$$= 26.46$$

$$\sigma_1{}^2 = \frac{172{,}772}{30} - \left(\frac{2{,}270}{30}\right)^2 = 33.62$$

$$R_{1\cdot 23} = \sqrt{1 - \frac{26.46}{33.62}} = 0.46$$

**102b. Partial correlation.**—For simple correlation we obtain two regression equations. These equations are of the form

$$y = r\frac{\sigma_y}{\sigma_x}x, x = r\frac{\sigma_y}{\sigma_x}y$$

If we use $x_1$ instead of $y$ and $x_2$ instead of $x$, these equations may be written

$$x_1 = b_{12}\,x_2, \quad x_2 = b_{21}\,x_1$$

We saw for this case that the simple correlation coefficient $r_{12}$ was given by the formula

$$r_{12} = r_{21} = \sqrt{b_{12} \cdot b_{21}}$$

where $b_{12}$ and $b_{21}$ were the coefficients of regression.

In the present instance, our regression equation is

$$X_1 = a + b_{12\cdot3}\,X_2 + b_{13\cdot2}\,X_3$$

where the coefficients $b_{12\cdot3}$ and $b_{13\cdot2}$ are called partial regression coefficients. The partial coefficient of correlation between $X_1$ and $X_2$, $X_3$ being held constant, is denoted by $r_{12\cdot3}$ and is given by the formula:

$$r_{12\cdot3} = \sqrt{b_{12\cdot3} \cdot b_{21\cdot3}}$$

For $n$ variables this would become

$$r_{12\cdot34\,\ldots\,n} = \sqrt{b_{12\cdot34\,\ldots\,n} \cdot b_{21\cdot34\,\ldots\,n}}$$

The coefficient of partial correlation between $X_1$ and $X_3$ is $r_{13\cdot2}$ where

$$r_{13\cdot2} = \sqrt{b_{13\cdot2}\,b_{31\cdot2}}$$

For $n$ variables this would become

$$r_{13\cdot2456\,\ldots\,n} = \sqrt{b_{13\cdot2456\,\ldots\,n}b_{31\cdot2456\,\ldots\,n}}$$

For the present illustrative problem we have

$$r_{12\cdot3} = \sqrt{0.111 \times 0.704} = 0.28$$
$$r_{13\cdot2} = \sqrt{0.188 \times 0.277} = 0.23$$
$$r_{23\cdot1} = \sqrt{0.121 \times 0.028} = 0.058$$

In obtaining this coefficient of partial correlation, we have secured a measure of the correlation which exists between two variables while the other variables are held constant.

Another method of obtaining this measure would be as follows: if our data were extensive enough, we might be able to pick out a number of pairs of values of $X_1$ and $X_2$ for which the corresponding pairs of values of $X_3$ were identical. A correlation coefficient between $X_1$ and $X_2$ for such a set of values would not be affected by fluctuations in $X_3$. Usually our data are too limited to enable us to proceed in this manner. In this case we use the method described above.

### Exercises.

1. Verify the computations in the illustrative problem.

2. Compute multiple and partial correlation coefficients for the following data (*taken from Colorado College examinations, September, 1925*):

| *Thorndike Intelligence* | *Freshman Mathematics* | *Freshman English* | *Thorndike Intelligence* | *Freshman Mathematics* | *Freshman English* |
|---|---|---|---|---|---|
| 63 | 85 | 75 | 49 | 65 | 78 |
| 45 | 72 | 75 | 75 | 82 | 72 |
| 59 | 85 | 85 | 59 | 40 | 75 |
| 68 | 82 | 90 | 80 | 61 | 78 |
| 49 | 40 | 76 | 46 | 85 | 78 |
| 52 | 85 | 82 | 49 | 40 | 40 |
| 70 | 80 | 86 | 79 | 85 | 76 |
| 59 | 83 | 81 | 55 | 40 | 40 |
| 53 | 40 | 75 | 66 | 75 | 85 |
| 48 | 55 | 75 | 68 | 95 | 91 |
| 66 | 77 | 87 | 65 | 83 | 77 |
| 79 | 93 | 87 | 53 | 73 | 72 |
| 71 | 76 | 63 | 51 | 35 | 35 |
| 69 | 75 | 95 | 70 | 88 | 85 |
| 45 | 65 | 70 | 69 | 75 | 80 |

### References.

Mills, F. C., "Statistical Methods," Henry Holt and Co., New York, 1924, Chapter XIV. Solution in detail of the relation between corn yield and temperature in Kansas for June, July, and August, 1890–1922.

Crum and Patton, "Economic Statistics," A. W. Shaw Co., New York, 1925, p. 259ff. Correlation between population, debits, and clearings, for thirty selected cities.

**103. Coefficient of contingency.**—We have shown how to obtain a measure of relationship for non-quantitative data classified so as to yield a four-fold table. We will now exhibit a method of obtaining a measure of relationship for non-quantitative data, for which the classification is finer.

Let us suppose that the classification has been such as to yield $R$ rows and $C$ columns of a correlation table. This table would then have $rc$ compartments or cells. Let us use the following notation:

$N$ = total frequency, or observations, or measures.
$f_r$ = number of measures in row number $r$.
$f_c$ = number of measures in column number $c$.
$n_{rc}$ = number of measures in compartment determined by the intersection of row $r$ and column $c$.

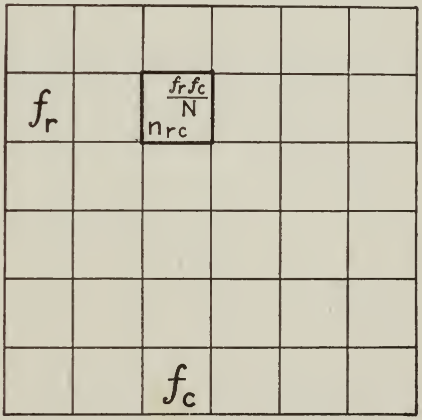


Fig. 44-A.

The probability that any one measure will fall somewhere in the row $r$ is $\frac{f_r}{N}$. The probability that any one measure will fall somewhere in the column $c$ is $\frac{f_c}{N}$. Hence, the probability, due to pure chance, that any one measure will fall in the cell at the intersection of this row and column is $\frac{f_r \cdot f_c}{N^2}$. Then, out of $N$ measures, the number of measures that one should expect to find in this cell would be $\frac{f_r f_c}{N}$. The number of measures actually in this cell is $n_{rc}$. The difference

$$n_{rc} - \frac{f_r f_c}{N}$$

must be due to some tendency for the two characters under consideration to be associated. These differences are squared to prevent cancellation of positive and negative items. Since it is only the relative size of these differences that is significant, we divide each squared difference by the frequency due to pure chance. What is called the *mean square contingency,* denoted by $\phi^2$, is then defined as follows:

$$\phi^2 = \frac{1}{N}\sum \frac{\left[n_{rc} - \frac{f_r f_c}{N}\right]^2}{\frac{f_r \cdot f_c}{N}}$$

the summation to extend to all compartments of the table.

Pearson's *coefficient C* of mean square contingency is

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

We have

$$\begin{aligned}
\phi^2 &= \frac{1}{N}\sum \frac{\left[n_{rc} - \frac{f_r f_c}{N}\right]^2}{\frac{f_r \cdot f_c}{N}} \\
&= \frac{1}{N}\sum \left[N\frac{n^2{}_{rc}}{f_r f_c} - 2n_{rc} + \frac{f_r f_c}{N}\right] \\
&= \sum \frac{n^2{}_{rc}}{f_r f_c} - \frac{1}{N}2\sum n_{rc} + \frac{1}{N^2}\sum f_r \sum f_c \\
&= \sum \frac{n^2{}_{rc}}{f_r f_c} - \frac{1}{N}\cdot 2N + \frac{1}{N^2}\cdot N \cdot N \\
&= \sum \frac{n^2{}_{rc}}{f_r f_c} - 1
\end{aligned}$$

For an $n$-fold table in which the measures occur only in the principal diagonal, we have

$$\begin{aligned}
\phi^2 &= \frac{a_1{}^2}{a_1{}^2} + \frac{a_2{}^2}{a_2{}^2} + \cdot \cdot \cdot + \frac{a_{n-1}{}^2}{a_{n-1}{}^2} + \frac{a_n{}^2}{a_n{}^2} - 1 \\
&= n - 1
\end{aligned}$$

| | | | | |
|---|---|---|---|---|
| $a_1$ | | | | |
| | $a_2$ | | | |
| | | $\cdot$<br>$\cdot$<br>$\cdot$ | | |
| | | | $a_{n-1}$ | |
| | | | | $a_n$ |

**Fig. 44.-B.**

Let us illustrate by means of the data in table 62 from an article by K. Pearson in *Biometrika,* vol. III, p. 182.

**Table 62.—Athletic Capacity.**

FIRST BROTHER

| SECOND BROTHER | | *Athletic* | *Betwixt* | *Non-Athletic* | TOTALS |
|---|---|---|---|---|---|
| | *Athletic* | 906 | 20 | 140 | 1,066 |
| | *Betwixt* | 20 | 76 | 9 | 105 |
| | *Non-Athletic* | 140 | 9 | 370 | 519 |
| | TOTALS | 1,066 | 105 | 519 | 1,690 |

$$\phi^2 = \frac{(906)^2}{(1{,}066)^2} + \frac{(20)^2}{(105)(1{,}066)} + \frac{(140)^2}{(519)(1{,}066)} + \frac{(20)^2}{(1{,}066)(105)} + \frac{(76)^2}{(105)^2}$$

$$+ \frac{9^2}{(519)(105)} + \frac{(140)^2}{(1{,}066)(519)} + \frac{9^2}{(105)(519)} + \frac{(370)^2}{(519)^2} - 1$$

$$= 0.8354$$

$$C = \sqrt{\frac{0.8354}{1 + 0.8354}} = 0.67$$

### Exercises.

For the following sets of data compute $\phi^2$ and $C$.

1. *H. Waite, Biometrika, vol. X, p. 472.* Correlation of finger markings

LEFT RING FINGER

|  |  | *Arch* | *Small Loop* | *Large Loop* | *Whorl* | *Composite* | TOTALS |
|---|---|---|---|---|---|---|---|
| LEFT LITTLE FINGER | *Arch*.............. | 17 | 13 | 4 | 0 | 1 | 35 |
|  | *Small loop*......... | 48 | 474 | 294 | 99 | 44 | 959 |
|  | *Large loop*.......... | 1 | 92 | 390 | 212 | 73 | 768 |
|  | *Whorl*............. | 0 | 3 | 12 | 120 | 15 | 150 |
|  | *Composite*.......... | 0 | 1 | 12 | 60 | 15 | 88 |
|  | TOTALS.......... | 66 | 583 | 712 | 491 | 148 | 2,000 |

$\phi^2 = 0.5047;\ C = 0.58$

2. *K. Pearson, Biometrika, vol. III, p. 190.* On Inheritance of Characters.

HANDWRITING
FIRST BROTHER

|  |  | *Very Good* | *Good* | *Moderate* | *Poor* | *Bad* | *Very Bad* | TOTALS |
|---|---|---|---|---|---|---|---|---|
| SECOND BROTHER | *Very Good*........... | 52 | 51 | 27.5 | 3 | 1 | . . | 134.5 |
|  | *Good*............... | 51 | 335 | 224.5 | 32 | 4 | 1 | 647.5 |
|  | *Moderate*............ | 27.5 | 224.5 | 406 | 101.5 | 15.5 | 2 | 777 |
|  | *Poor*................ | 3 | 32 | 101.5 | 96 | 15 | 2 | 249.5 |
|  | *Bad*................ | 1 | 4 | 15.5 | 15 | 7 | 1 | 43.5 |
|  | *Very Bad*........... | ..... | 1 | 2 | 2 | 1 | 4 | 10 |
|  | TOTALS........... | 134.5 | 647.5 | 777 | 249.5 | 43.5 | 10 | 1,862 |

## Chapter X

## INDEX NUMBERS

**104. Index numbers.**—An index number is a statistical device used to express the average change in the magnitude of a group of related variables. It is a representative number. With respect to prices, an index number represents an average change of prices from one point of time to another.

There would be no need of an index number if the prices of different commodities rose and fell in perfect unison. But prices do not rise and fall in unison. They seem to scatter and disperse; some rise while others fall. But there is a definite average movement of the scattering prices. This average is the index number. In its simplest form it is a number which expresses the relation which the price of a given commodity at any time bears to its price at some fixed time. To illustrate, the average farm price of corn on Dec. 1, 1899, was 30.3 cents and on Dec. 1, 1901, was 60.5 cents. Then the price in 1901 relative to the price in 1899 as a base is represented by the number 2. This number 2 is the index number of price in 1901 relative to 1899 as a base. If we desire to express results as percentages, we use 200 as the index number of price in place of 2.

Table 63 gives the index number of the average farm price of corn, Dec. 1, with respect to 1899 as a base. (See p. 180.)

The changes in price of a single commodity or a group of commodities is the net resultant of all forces influencing such changes. These forces are seasonal and climatic influences, reductions and increases in cost, changes in the technique of production, changes in style and habit, the introduction of substitutes, the increased intensity of competition, development of favorable or unfavorable legislation, personal salesmanship, and so forth. These forces are subject to the general influence of monetary inflation or deflation. An index number is a number which measures the net resultant of all forces at work. By means of this index number we are able to discover and study the trend of prices for raw materials, stocks, bonds, money, labor, transportation, etc.

**Table 63.—Index Numbers of the Average Farm Price of Corn, Dec. 1, Base 1899.**

| *Year* | *Price* | *Index* | *Year* | *Price* | *Index* |
|---|---|---|---|---|---|
| 1899 | 30.3 | 1.00 | 1905 | 41.2 | 1.36 |
| 1900 | 35.7 | 1.18 | 1906 | 39.9 | 1.32 |
| 1901 | 60.5 | 2.00 | 1907 | 51.6 | 1.70 |
| 1902 | 40.3 | 1.33 | 1908 | 60.6 | 2.00 |
| 1903 | 42.5 | 1.40 | 1909 | 57.9 | 1.91 |
| 1904 | 44.1 | 1.45 | 1910 | 48.0 | 1.58 |

Index numbers are computed for many things besides prices. They are used in the study of changes in production and consumption of goods, to measure changes in volume of trade, in the study of deferred payments, the supply of money and credit, to regulate employment and rates of wages, and to compare real incomes at different times and places.

The purpose of index numbers may be general or specific in nature, but, when used properly, they serve as guides to point out the safest path along which modern business should travel.

**105. Composite index numbers.**—A composite index number is based upon more than one commodity. These index numbers are classified as *aggregative* or *relative.* An aggregative index number is based upon the sums of actual prices or total values. A relative index number is computed by finding a simple relative index number for each commodity and then taking some kind of an average of these relatives. Index numbers may again be classified as *weighted* or *unweighted.*

## A. Aggregative Index Numbers.

**106. Aggregates of actual prices.**—Let us illustrate this method by obtaining an index number for the price of corn, wheat, and oats. Data are from the Yearbook of the Department of Agriculture and give the average farm price in cents on Dec. 1.

**Table 64.—Aggregative Index Number of Actual Prices of Corn, Wheat, and Oats. Base, 1899.**

| *Year* | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | *Corn* | *Wheat* | *Oats* | $\Sigma p_i$ | $\Sigma p_i/\Sigma p_o$ |
| 1899 | 30.3 | 58.4 | 31.3 | 120.0 | 1.00 |
| 1900 | 35.7 | 61.9 | 29.1 | 126.7 | 1.06 |
| 1901 | 60.5 | 62.4 | 31.7 | 154.6 | 1.29 |
| 1902 | 40.3 | 63.0 | 44.3 | 147.6 | 1.23 |
| 1903 | 42.5 | 69.5 | 47.2 | 159.2 | 1.33 |
| 1904 | 44.1 | 92.4 | 40.2 | 176.7 | 1.47 |

The index numbers in column four are obtained by adding together the prices which appear in any given row. Thus

$$30.3 + 58.4 + 31.3 = 120.0$$

It is the custom to express this index number as follows:

$$\text{(1)} \qquad \text{Aggregative index of prices} = \Sigma p_i$$

If we use 1899 as a base year, we obtain the numbers in column five. These numbers are the ratios of the actual sum of prices in the given year to the actual sum of prices in the base year. Thus, taking 1900 as the given year and 1899 as the base year, we have

$$\frac{35.7 + 61.9 + 29.1}{30.3 + 58.4 + 31.3} = \frac{126.7}{120.0} = 1.06$$

The general formula for this index number, which is a relative of the actual prices, is

$$\text{(2)} \qquad \frac{\Sigma p_i}{\Sigma p_o}$$

where $\Sigma p_o$ is the sum of prices in the base year, and $\Sigma p_i$ is the sum of prices in the given year.

**107. Weighted aggregates of actual prices.**—There is no index number which is truly unweighted. When no weights are specified, the weights are unity.

The average workingman's family spends about ten times as much for milk as it does for bacon. If milk is 10 cents a quart, or 5 cents a pound, and bacon is 30 cents a pound, undue importance is given to fluctuations in the price of bacon when an index number is constructed on the basis of aggregates of actual prices. A more representative index number is one in which the prices are weighted or multiplied by the quantity used, giving the total value of the product used in a given period. The weights are usually the quantities $q_o$ used or produced in the base year.

In obtaining general expressions for index numbers, we make use of the customary notation which is explained in the following paragraph.

If

$$p_i' = \text{price of first commodity in year } i$$

$$p_i'' = \text{price of second commodity in year } i$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$p_i^{(n)} = \text{price of } n\text{th commodity in year } i,$$

then the total value[1] of the commodities under consideration in the $i$-th year after the base year, is given by

$$p_i'q_o' + p_i''q_o'' + \cdot \cdot \cdot + p_i^{(n)}q_o^{(n)}$$

This sum is usually represented by the symbol

$$\sum_{1}^{n}{}_j p_i^{(j)}q_o^{(j)} \tag{3}$$

or sometimes more simply by $\Sigma p_i q_o$ or $\Sigma p_1 q_o$.

[1] This is not strictly true. Actually the total value is

$$p_i'q_i' + p_i''q_i'' + \cdot \cdot \cdot + p_i^{(n)}q_i^{(n)}$$

However, the value is commonly representea by the expression given in the text.

It is desirable at times to express these weighted aggregates as relatives of the total value in the base year. Our index number is then represented by

$$\text{(4)} \qquad \frac{\Sigma p_1 q_o}{\Sigma p_o q_o} = \frac{\Sigma p_o q_o \dfrac{p_1}{p_o}}{\Sigma p_o q_o}$$

This is perhaps the most generally useful index number. The second form of its expression shows that it is an arithmetic average of relative prices.

**Table 65.—Average Price and Quantity of the Various Units Used by the Workingman's Family.**

| Units | *Sirloin Steak* | *Round Steak* | *Bacon* | *Eggs* | *Butter* | *Milk* | *Flour* | *Potatoes* | *Sugar* |
|---|---|---|---|---|---|---|---|---|---|
| | Lb. | Lb. | Lb. | Doz. | Lb. | Qt. | ⅛ Bbl. | Peck | Lb. |
| Weights | 15 | 40 | 13 | 70 | 76 | 424 | 8 | 50 | 145 |
| *1913* | $.254 | $.223 | $.270 | .345 | .383 | .089 | .809 | .255 | .055 |
| *14* | .259 | .236 | .275 | .353 | .362 | .089 | .833 | .270 | .059 |
| *15* | .257 | .230 | .269 | .341 | .358 | .088 | 1.029 | .225 | .066 |
| *16* | .273 | .245 | .287 | .375 | .394 | .091 | 1.078 | .405 | .080 |
| *17* | .315 | .290 | .410 | .481 | .487 | .112 | 1.715 | .645 | .093 |
| *1918* | .389 | .369 | .529 | .569 | .577 | .139 | 1.642 | .480 | .097 |
| *19* | .417 | .389 | .554 | .628 | .678 | .155 | 1.764 | .570 | .113 |
| *20* | .437 | .395 | .523 | .681 | .701 | .167 | 1.985 | .945 | .194 |
| *21* | .388 | .344 | .427 | .509 | .517 | .146 | 1.421 | .465 | .080 |
| *22* | .374 | .323 | .398 | .444 | .479 | .131 | 1.250 | .420 | .073 |

Let us use the data in table 65 to construct an index number of a weighted aggregate of actual prices, using 1913 as the base year and 1922 as the given year. The weights used are adapted from those given in the Bureau of Labor Statistics Bulletin 357, pp. 108–109, and represent the quantities in the specified units which are used per year by the average workingman's family.

Thus, we find that the index number for 1913 is 134.431, while that for 1922 is 188.317.

If we desire to express the aggregate of prices in 1922 relative to 1913 as a base, we use formula (4) and find

$$\left.\begin{array}{c}\text{Weighted index number of prices in}\\ \text{1922 with prices in 1913 as a base}\end{array}\right\} = \frac{\Sigma p_1 q_o}{\Sigma p_o q_o} = \frac{188.317}{134.431} = 1.40$$

**Table 66.—Computation of Index Number as Weighted Aggregate of Actual Prices.**

| | *Quantity Used in* 1913 | PRICE | | | |
|---|---|---|---|---|---|
| | | 1913 | 1922 | | |
| COMMODITY | $q_o$ | $p_o$ | $p_1$ | $p_o q_o$ | $p_1 q_o$ |
| Sirloin steak............ | 15 lb. | $.254 | .374 | 3.810 | 5.610 |
| Round steak............ | 40 lb. | .223 | .323 | 8.920 | 12.920 |
| Bacon.................. | 13 lb. | .270 | .398 | 3.510 | 5.174 |
| Eggs................... | 70 doz. | .345 | .444 | 24.150 | 31.080 |
| Butter................. | 76 lb. | .383 | .479 | 29.108 | 36.404 |
| Milk................... | 424 qts. | .089 | .131 | 37.736 | 55.544 |
| Flour.................. | 1 bbl. | 6.472 | 10.000 | 6.472 | 10.000 |
| Potatoes............... | 50 pk. | .255 | .420 | 12.750 | 21.000 |
| Sugar.................. | 145 lb. | .055 | .073 | 7.975 | 10.585 |
| | | | TOTALS... | 134.431 | 188.317 |

Expressed as percentages, the index number for 1913 is 100, while that for 1922 is 140.

### Exercise.

Compute by means of formula (4) the index of prices for the other years in table 65, using 1913 as a base with an index number of 100.

## B. Relative Index Numbers.

**108. Serial relative.**—Compute for each article the ratio of the price for each year to the price for some fixed base year. Then an index number for any desired year can be obtained by taking some one of the various averages of these relatives for that year. We illustrate this method below for the unweighted arithmetic average, median, mode, and geometric average. For more detailed discussion of these averages and for a discussion of other averages the reader is referred to Prof. Irving Fisher's treatise, "The Making of Index Numbers."

Table 67.—Computation of Serial Relative for 1922 with 1913 as Base for Data, Table 65.

| | | 1913 | 1922 | Serial Relative |
|---|---|---|---|---|
| Commodity | *Unit* | $p_o$ | $p_1$ | $p_1/p_o$ |
| Sirloin steak | lb. | $.254 | .374 | 1.47 |
| Round steak | lb. | .223 | .323 | 1.45 |
| Bacon | lb. | .270 | .398 | 1.47 |
| Eggs | doz | .345 | .444 | 1.29 |
| Butter | lb. | .383 | .479 | 1.25 |
| Milk | qt. | .089 | .131 | 1.47 |
| Flour | bbl. | 6.472 | 10.000 | 1.55 |
| Potatoes | pk. | .255 | .420 | 1.65 |
| Sugar | lb. | .055 | .073 | 1.33 |
| | | | | 12.95 |

**Exercise.**

Compute from table 65 the serial relative for the other years.

**109. Arithmetic average of relatives.**—Take the sum of the relatives $p_1/p_o$ for a given year and divide by the number $N$ of articles. We have

$$\text{Index number: Arith. Av. of relatives} = \frac{\Sigma\left(\frac{p_1}{p_o}\right)}{N}$$

For the list of nine articles in table 67, we find that an index number on this basis for the year 1922 with 1913 as base is as follows:

$$\text{Index number} = \frac{12.95}{9} = 1.44$$

An arithmetic average of relative prices, weighted according to the total values in the base year, is always equal to a relative of weighted aggregates of actual prices constructed from the same data. Thus

$$\frac{\frac{p_1'}{p_o'} \times p_o'q_o' + \frac{p_1''}{p_o''} p_o''q_o'' + \frac{p_1'''}{p_o'''} p_o'''q_o''' + \cdots}{p_o'q_o' + p_o''q_o'' + p_o'''q_o''' + \cdots} = \frac{\Sigma p_1q_o}{\Sigma p_oq_o}$$

**Exercise.**

Compute from table 65 the arithmetic average of relatives for the other years.

**110a. Median of relatives.**—The arithmetic average of relatives is sensitive to changes in all of the items. Special circumstances might cause a large variation in one item while the other items remained unaffected. For example, in obtaining an index of living costs, the price of rice might double, and consequently less would be used; in its place more potatoes might be used with but little variation in the price of potatoes. As a net result, the cost of living would be changed but little. But if an index is computed on a basis of an arithmetic average of relatives, this index would show considerable fluctuation, for no account is taken of the diminished consumption of an item with a relatively high price. A median of relatives is a better average to use. It is not so sensitive to great variations in relative prices of a few articles.

To find the median relative, arrange the relatives in a frequency table and find the median by the usual process. For the data in table 67, we find the index number on the basis of the median relative to be 1.47.

Index number: median relative = 1.47

**110b. Mode of relatives.**—Arranging the relatives in table 67 in a frequency table and determining the mode, we find for an index number the following:

Index number: mode of relatives = 1.47

**111. Geometric average of relatives.**—Expressed as relatives, the price in the base year is frequently represented as 100. The range of the relatives in one direction is limited to the range from 0 to 100. In the other direction the range is unlimited. In practice it has been found that, if one plots a frequency curve of relative prices, one obtains an asymmetrical curve with the longest tail in the direction of the unlimited range. Furthermore, it has been found that, if one takes the logarithms of the relative prices and constructs a frequency curve, this curve is a fairly symmetrical bell-shaped curve. For a symmetrical distribution the arithmetic average, mode, and median coincide. Then one of these averages is suggested for the logarithms of the relative frequencies. But the arithmetic average of the logarithms of a set of numbers is equal to the logarithm of the geometric average of the numbers themselves.

The formula for the unweighted geometric average of $N$ relatives is

$$G = \sqrt[N]{\frac{p_1'}{p_o'} \cdot \frac{p_1''}{p_o''} \cdot \frac{p_1'''}{p_o'''} \cdots \frac{p_1^{(n)}}{p_o^{(n)}}}$$

The principal objection to this index number is the tediousness of the necessary computations. Hence, it is seldom used. For the data in table 67, we have

$$G = \sqrt{(1.47)^3 \times 1.45 \times 1.29 \times 1.25 \times 1.55 \times 1.65 \times 1.33} = 1.435$$

**112. Fisher's ideal index.**—An index number computed from a weighted aggregate of actual prices, as outlined in article 107, fails to take account of the fact that the quantities of the various items used vary from year to year. Prof. Fisher has constructed an index number which takes into account both the varying price and the varying quantity. This ideal index is as follows:

$$\text{Ideal index} = \sqrt{\frac{\Sigma p_1 q_o}{\Sigma p_o q_o} \frac{\Sigma p_1 q_1}{\Sigma p_o q_1}}$$

There is a difficulty in the use of this formula in that statistics are not available for the year to year variations of the quantities of the various items. The necessary computations are also tedious.

Prof. Fisher has offered a substitute formula:

$$\frac{\Sigma(q_o + q_1)p_1}{\Sigma(q_o + q_1)p_o} \qquad \textit{(Fisher's formula 2,153)}$$

for which the time of calculation is materially reduced. We will refer to this formula as Fisher's modified index. This index has the same difficulty as the ideal index with respect to the inability to obtain the yearly or monthly values for

$q_1$. This formula has been approved[2] by Edgeworth, Fisher, Marshall, and Walsh.

Formula[3] 2,153 will, under all ordinary circumstances, be sufficiently close to Formula 353 (Fisher's Ideal) to serve as a short cut substitute.

Taking[4] into account accuracy, speed, ease of manipulation and intelligibility, Formula 2,153 seems, on the whole, to take the highest rank for ordinary practical use.

### Exercises.

Compute for the data in table 65 an index number for the various years by the method of

(a) median of relatives; (b) mode of relatives;
(c) geometric average of relatives; (d) ideal index;
(e) Fisher's modified formula.

## C. Time and Factor Reversal Tests.[5]

**113. Time reversal test.**—A good index number should be one such that, if the base year is shifted, the relative size of the index numbers is not changed. For example, if an index number for 1910 with 1899 as base is 2, then the index number for 1899 with 1910 as base should be $\frac{1}{2}$. That is, the product of the two index numbers should be unity.

Let us apply this test to some of the index numbers previously discussed. Let us use $I_o$ to represent the index number of any second year with respect to a first year as a base; and $I_1$ to represent the index number of the first year with respect to the second year as a base.

(a) *Simple aggregate of actual prices expressed as relatives.*—For this index number we have

$$I_o = \frac{\Sigma p_1}{\Sigma p_o}, \quad I_1 = \frac{\Sigma p_o}{\Sigma p_1}$$

[2] Fisher, Irving, "The Making of Index Numbers," p. 484, No. 2,153. Third edition, revised, Houghton Mifflin Company, Boston, 1927.

[3] Fisher, loc. cit., p. 428. See also pp. 329 and 247.

[4] Fisher, loc. cit., p. 349.

[5] For a discussion of the so-called *circular test* consult Fisher, "The Making of Index Numbers," p. 270 ff. Fisher states that "the circular test is theoretically a mistaken one."

Whence

$$I_oI_1 = 1$$

Thus, we see that this index number meets the time reversal test.

Using the data in article 106, we have for the years 1899 and 1900,

$$I_o = \frac{126.7}{120.0}, \quad I_1 = \frac{120.0}{126.7}$$

Whence

$$I_oI_1 = \frac{126.7}{120.0} \cdot \frac{120.0}{126.7} = 1$$

(b) *Weighted aggregates of actual prices expressed as relatives.*—For this index number we have

$$I_o = \frac{\Sigma p_1q_o}{\Sigma p_oq_o}, \quad I_1 = \frac{\Sigma p_oq_1}{\Sigma p_1q_1}$$

Whence

$$I_0I_1 \neq 1$$

This index number does not meet the time reversal test. To illustrate, let us use the data in table 68, taken from the Yearbook of the Department of Agriculture for 1925.

**Table 68.—Data Used in Time Reversal Test for Weighted Aggregates of Actual Prices Expressed as Relatives.**

| | 1913 | | | 1925 | | |
|---|---|---|---|---|---|---|
| | $q_o$ | $p_o$ | $p_oq_o$ | $q_1$ | $p_1$ | $p_1q_1$ |
| | *Production 1,000 Bu,* | *Farm Price per Bu. Dec. 1* | *Farm Value $* | *Production 1,000 Bu.* | *Farm Price per Bu. Dec. 1* | *Farm Value $* |
| Corn.................... | 2.446,988 | 69.1¢ | $1,690,869 | 2,900,581 | 67.4¢ | $1,954,992 |
| Wheat.................... | 763,380 | 79.9 | 610,122 | 669,365 | 141.6 | 947,993 |
| | $\Sigma p_oq_o$ = $2,300,991 | | | $\Sigma p_1q_1$ = $2,902,985 | | |

$$\begin{aligned}
\Sigma p_oq_1 &= 2{,}900{,}581 \times 0.691 + 669{,}365 \times 0.799 \\
&= 2{,}004{,}301.471 + 534{,}822.635 \\
&= 2{,}539{,}124.106 \\
\Sigma p_1q_o &= 2{,}446{,}988 \times 0.674 + 763{,}380 \times 1.416 \\
&= 1{,}649{,}269.912 + 1{,}080{,}946.080 \\
&= 2{,}730{,}215.992
\end{aligned}$$

Whence

$$I_o = \frac{2{,}730{,}215.99}{2{,}300{,}991.00}, \quad I_1 = \frac{2{,}539{,}124.106}{2{,}902{,}985.00}$$

$$I_oI_1 = \frac{2{,}730{,}215.99}{2{,}300{,}991.00} \times \frac{2{,}539{,}124.106}{2{,}902{,}985.00} = 1.038 \neq 1$$

(c) *Median of relatives.*—The relatives of the prices in year one with respect to year two as a base are the reciprocals of the relatives of the prices in year two with respect to year one as a base. Hence, if the numbers in both series are arranged in order of magnitude, the median of the one series is the reciprocal of the median in the other series, since if

$$a < b < c < \cdots$$

then

$$\frac{1}{a} > \frac{1}{b} > \frac{1}{c} > \cdots$$

Since the product of a number and its reciprocal is unity, we see that this index number meets the time reversal test.

To illustrate, let us use the data in table 69, taken from the Yearbook of the Department of Agriculture, 1925.

**Table 69.—Data Used in Time Reversal Test for Median of Relatives.**

| (1) | (2) | (3) | Relative | |
|---|---|---|---|---|
| | *Price in ¢ per Bu.* | | *Base* | |
| | 1913 | 1925 | 1913 | 1925 |
| Corn | 69.1 | 67.4 | 0.975 | 1.025 |
| Wheat | 79.9 | 141.6 | 1.772 | 0.564 |
| Oats | 39.2 | 38.1 | 0.972 | 1.029 |
| Rye | 63.4 | 78.1 | 1.232 | 0.812 |
| Flaxseed | 119.9 | 226.5 | 1.889 | 0.529 |
| Barley | 53.7 | 58.6 | 1.091 | 0.916 |
| Buckwheat | 75.5 | 89.2 | 1.181 | 0.846 |

The median of relative prices with 1913 as a base is 1.181. The median of relative prices with 1925 as a base is 0.846. We have

$$1.181 \times 0.846 = 0.999$$

More accurately, using the data in columns (2) and (3) for buckwheat,

$$\frac{89.2}{75.5} \times \frac{75.5}{89.2} = 1$$

(d) *Mode of relatives.*—The relatives of the one series are the reciprocals of the relatives of the other series. If a number has a greatest frequency in the one series, its reciprocal will have the greatest frequency in the other series. The product of a number and its reciprocal is unity. Hence, this index number meets the time reversal test.

(e) *Geometric average of relatives.*—For the geometric average of relatives, we have

$$I_o = \sqrt[N]{\frac{p_1'}{p_o'}\,\frac{p_1''}{p_o''}\,\frac{p_1'''}{p_o''}\cdots}, \quad I_1 = \sqrt[N]{\frac{p_o'}{p_1'}\,\frac{p_o''}{p_1''}\,\frac{p_o'''}{p_1'''}\cdots}$$

Whence

$$I_oI_1 = 1$$

Hence, this index number meets the time reversal test.

### Exercises.

Determine whether the following index numbers meet the time reversal test:

| | *Ans.* |
|---|---|
| 1. Fisher's ideal index. | Yes. |
| 2. Fisher's modified index. | Yes. |
| 3. Arithmetic average of relatives. | No. |
| 4. Harmonic average of relatives. | No. |

**114. Factor reversal test.**—Total value is the product of price and quantity. A good index number should be such that if separate indices are computed for price and for quantity, their product will give a correct index of total value.

Let us use $P$ to indicate index of price, $Q$ to indicate index of quantity, and $V$ to indicate index of value.

(a) *Fisher's ideal index.*—Fisher's ideal index for price is

$$P = \sqrt{\frac{\Sigma p_1q_o}{\Sigma p_oq_o} \cdot \frac{\Sigma p_1q_1}{\Sigma p_oq_1}}$$

The prices are weighted in the first factor by the quantities in the base year, in the second factor by the quantities in the given year.

An index number for quantity, constructed on the same basis, would have the quantities in the first factor weighted with the prices in the base year, the quantities in the second factor weighted with the prices in the given year. Thus, we would have

$$Q = \sqrt{\frac{\Sigma q_1 p_o}{\Sigma q_o p_o} \cdot \frac{\Sigma q_1 p_1}{\Sigma q_o p_1}}$$

Whence

$$P \times Q = \frac{\Sigma p_1 q_1}{\Sigma p_o q_o} = V$$

Thus, this index number satisfies the factor reversal test.

(b) *Weighted aggregates of actual prices expressed as relatives.*—For this index, we have

$$P = \frac{\Sigma p_1 q_o}{\Sigma p_o q_o}, \quad Q = \frac{\Sigma q_1 p_o}{\Sigma q_o p_o}$$

Whence

$$P \times Q \neq V = \frac{\Sigma p_1 q_1}{\Sigma p_o q_o}$$

Thus, this index number does not satisfy the factor reversal test.

### Exercises.

Determine whether the following index numbers meet the factor reversal test.

*Ans.*

1. Relatives of unweighted aggregates of actual prices. No.
2. Median of relative prices. No.
3. Mode of relative prices. No.
4. Geometric average of relative prices. No.
5. Fisher's modification of the ideal index. No.

## D. Bias.

**115. Bias.**[6]—One requisite of a good index number is that it shall satisfy the time reversal test, which requires that the product of the index number for any given year on a base year and the index number for the base year on the given year shall be unity.

[6] See article on *Index Number Bias* by W. V. Lovitt, *Journal of the American Statistical Association,* March, 1928, pp. 10–17.

If this product is greater than unity, the index number is said to have an upward bias. If this product is less than unity, the index number is said to have a downward bias.

We have seen that the geometric average of relatives satisfies the time reversal test and thus has no bias. The arithmetic and harmonic averages of relatives do not satisfy the time reversal test and thus have what we call *type* bias.

Let us illustrate by means of the data in table 70, taken from the Yearbook of the Department of Agriculture for 1925.

**Table 70.—Price and Relative Price of Corn, Wheat, and Flaxseed for 1910–1911.**

| | Price in ¢ per Bu. | | Relative Base | | Reciprocals of Relatives | |
|---|---|---|---|---|---|---|
| | *1910* | *1911* | *1910* | *1911* | *1910* | *1911* |
| Corn | 48.0 | 61.8 | 128.8 | 77.6 | 0.00776 | 0.01288 |
| Wheat | 88.1 | 88.0 | 99.0 | 101.0 | 0.01070 | 0.00990 |
| Flaxseed | 231.7 | 182.1 | 78.7 | 127.1 | 0.01271 | 0.00787 |
| Aggregates | 367.8 | 331.9 | 306.5 | 305.7 | | |

Using the proper formula in each case, the numbers in table 71 are computed from those in table 70.

**Table 71.—Table Illustrating the Presence of Type Bias, Based on Taɒle 70.**

| Type | Index Number | | | |
|---|---|---|---|---|
| | *Base* | | *Product* | *Bias* |
| | *1910* | *1911* | | |
| Aggregative | 90.2 | 110.8 | 1.00 | None |
| Arithmetic average | 102.2 | 101.9 | 1.04 | Upward |
| Harmonic average | 98.1 | 97.9 | 0.96 | Downward |
| Median | 99.0 | 101.0 | 1.00 | None |
| Geometric average | 101.12 | 99.87 | 1.00 | None |

That the bias given in table 71 is always present is shown as follows:

For the unweighted arithmetic average of relatives, we have to show that

$$N^2 < \sum_{1}^{N}{}_{i} \frac{p_1{}^i}{p_o{}^i} \sum_{1}^{N}{}_{j} \frac{p_o{}^j}{p_1{}^j}$$

This inequality readily reduces to

$$O < \sum_{1}^{N}{}_{i} \sum_{1}^{N}{}_{j} (p_o{}^i p_1{}^j - p_o{}^j p_1{}^i)^2 \prod_{1}^{N}{}_{s} p_o{}^s \prod_{1}^{N}{}_{s} p_1{}^s (i \neq j;\ s \neq i, j)$$

This establishes the upward bias of the unweighted arithmetic average of relatives.

For the unweighted harmonic average of relatives, we have to show that

$$\frac{N}{\sum_{1}^{N}{}_{j} \frac{p_o{}^j}{p_1{}^j}} \frac{N}{\sum_{1}^{N}{}_{i} \frac{p_1{}^i}{p_o{}^i}} < 1$$

or

$$N^2 < \sum_{1}^{N}{}_{j} \frac{p_o{}^j}{p_1{}^j} \sum_{1}^{N}{}_{i} \frac{p_1{}^i}{p_o{}^i}$$

This is the same inequality that we met with in discussing above the unweighted arithmetic average of relatives. Thus, we have established the permanent downward bias of the unweighted harmonic average of relatives.

Fisher[7] outlines the four following methods of weighting by values:

I. Each weight = base year weight × base year quantity ($p_oq_o$)
II. Each weight = base year price × given year quantity ($p_oq_1$)
III. Each weight = base year price × base year quantity ($p_1q_o$)
IV. Each weight = given year price × given year quantity ($p_1q_1$)

Let us use these quantities as weights in a weighted arithmetic average of relatives.

Using I and applying the time reversal test, we have to discover whether

$$A \equiv \frac{\Sigma p_1q_o}{\Sigma p_oq_o} \cdot \frac{\Sigma p_oq_1}{\Sigma p_1q_1} \overset{>}{\underset{<}{=}} 1$$

[7] "The Making of Index Numbers," p. 54.

For the set of values

($\alpha$)

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 1 | 2 | 3 | 1 |
| 1 | 3 | 2 | 2 |

we find $A = \frac{36}{35} > 1$

For the set of values

($\beta$)

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 1 | 2 | 3 | 1 |
| 1 | 3 | 2 | 1 |

we find $A = \frac{24}{25} < 1$

For the set of values

($\gamma$)

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 1 | 2 | 3 | 1 |
| 1 | 3 | 2 | 1.5 |

we find $A = 1$

Thus, this index number has no permanent bias.

Using II and applying the time reversal test, we have to discover whether

$$B \equiv \frac{\Sigma p_o q_1 \frac{p_1}{p_o}}{\Sigma p_o q_1}\;\frac{\Sigma p_1 q_o \frac{p_o}{p_1}}{\Sigma p_1 q_o} = \frac{\Sigma p_1 q_1}{\Sigma p_o q_1}\;\frac{\Sigma p_o q_o}{\Sigma p_1 q_o} \substack{>\\=\\<} 1$$

**Exercise.**

Let the student show that for the set of values

($\alpha$) we have $B = \frac{35}{36} < 1$

($\beta$) we have $B = \frac{25}{24} > 1$

($\gamma$) we have $B = 1$

Thus, this index number has no fixed bias.

Using III, and applying the time reversal test, we have to discover whether

$$C \equiv \frac{\Sigma p_1 q_o \frac{p_1}{p_o}}{\Sigma p_1 q_o}\;\frac{\Sigma p_o q_1 \frac{p_o}{p_1}}{\Sigma p_o q_1} \substack{<\\=\\>} 1$$

The reader may verify that for the set of values ($\beta$) we have

$$C = \frac{25}{24} > 1.$$

For the set of values

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 1 | 2 | 3 | 2 |
| 1 | 3 | 2 | 1 |

we have $C = \frac{35}{36} < 1$

For the set of values

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 1 | 2 | 3 | 1.5 |
| 1 | 3 | 2 | 1 |

we have $C = 1$

Thus, this index number has no fixed bias.

Using IV and applying the time reversal test, we have to discover whether

$$D \equiv \frac{\Sigma p_1 q_1 \frac{p_1}{p_o}}{\Sigma p_1 q_1} \frac{\Sigma p_o q_o \frac{p_o}{p_1}}{\Sigma p_o q_o} \lesseqgtr 1$$

For the set of values

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 2 | 2 | 3 | 1 |
| 2 | 3 | 2 | 4 |

we find $D = \frac{65}{66} < 1$

For the set of values

| $p_o$ | $q_o$ | $p_1$ | $q_1$ |
|---|---|---|---|
| 2 | 2 | 3 | 1 |
| 2 | 3 | 2 | 3 |

we find $D = \frac{91}{90} > 1$

Thus, this index number has no fixed bias.

**116. Comparison of weighted arithmetic average of relatives using different weights.**—It is of interest to compare the relative size of index numbers computed on the basis of a weighted arithmetic average of relatives when I, II, III, and IV are used as weights.

*I versus III.*—If an index number computed on the basis of a weighted arithmetic average of relatives is always greater when $p_1q_o$ are used as weights than when $p_oq_o$ are used as weights, we have to show that

$$\frac{\sum_i p_o{}^i q_o{}^i \frac{p_1{}^i}{p_o{}^i}}{\sum_i p_o{}^i q_o{}^i} < \frac{\sum_j p_1{}^j q_o{}^j \frac{p_1{}^j}{p_o{}^j}}{\sum_j p_1{}^j q_o{}^j}$$

This inequality reduces to

$$0 < \sum_{1}^{N}{}_i \sum_{1}^{N}{}_j q_o{}^i q_o{}^j [p_o{}^i p_1{}^j - p_o{}^j p_1{}^i]^2 \prod_{1}^{N}{}_s p_o{}^s (s \neq i, j)$$

This inequality holds, for the right hand side, being the sum of a number of positive terms, is greater than zero. Thus, we see that I $<$ III with the single exception when the percentage increase (decrease) is the same for every commodity, when the two are equal.

*II versus IV.*—If an index number computed on the basis of a weighted arithmetic average of relatives is always greater when $p_1q_1$ are used as weights, than when $p_oq_1$ are used, then we have to show that

$$\frac{\Sigma p_oq_1\frac{p_1}{p_o}}{\Sigma p_oq_1} < \frac{\Sigma p_1q_1\frac{p_1}{p_o}}{\Sigma p_1q_1}$$

This inequality reduces to

$$0 < \Sigma q_1{}^iq_1{}^j(p_o{}^ip_1{}^j - p_o{}^jp_1{}^i)^2\prod_{1}^{N}{}_s\, p_o{}^s(s \neq i, j)$$

This inequality shows that II $<$ IV with the single exception when the percentage increase (decrease) is the same for every commodity, when the two are equal.

## Exercises.

But

$$\text{I} = \frac{\Sigma p_1q_o}{\Sigma p_oq_o}, \qquad \text{II} = \frac{\Sigma p_1q_1}{\Sigma p_oq_1},$$

$$\text{III} = \frac{\Sigma p_1q_o\frac{p_1}{p_o}}{\Sigma p_1q_o}, \quad \text{IV} = \frac{\Sigma p_1q_1\frac{p_1}{p_o}}{\Sigma p_1q_1}$$

Show that[8]

1. For the set of values ($\alpha$), I $>$ II
2. For the set of values ($\beta$), I $<$ II
3. For the set of values ($\alpha$), III $>$ IV
4. For the set of values ($\beta$), III $<$ IV
5. For the set of values ($\gamma$), III $=$ IV
6. For the set of values ($\alpha$), II $<$ III
7. For the set of values ($\beta$), II $=$ III
8. For the set of values

| $p_o$ | $q_o$ | $p_1$ | $q_1$ | |
|---|---|---|---|---|
| 1 | 2 | 3 | 2 | we have II $>$ III |
| 1 | 3 | 2 | 1 | |

[8] See article on *Index Number Bias* by W. V. Lovitt, *Journal of the American Statistical Association*, March, 1928.

9. For the set of values ($\alpha$), I < IV

10. For the set of values

| $p_0$ | $q_0$ | $p_1$ | $q_1$ | |
|---|---|---|---|---|
| 1 | 2 | 3 | 1 | we have I > IV |
| 1 | 3 | 2 | 3 | |

## E. American Index Numbers.

**117. Bureau of labor statistics.**—The best index number of wholesale prices in the United States is that published by the Bureau of Labor Statistics in the *Monthly Labor Review.* This index number is a weighted aggregate of actual prices of 404 commodities, with 1913 used as a base. The weights used are the quantities marketed in 1919. This index number represents the cost at wholesale of a specified bill of goods. The Bureau computes index numbers for nine commodity sub-groups, namely:

Building materials.
Chemicals and drugs.
Cloths and clothing.
Farm products.
Foods.
Fuel and lighting.
House furnishings.
Metals and metal products.
Miscellaneous.

Prices used are average monthly prices.

The Bureau of Labor Statistics publishes in the *Monthly Labor Review* an index number which may be designated as an index number of retail food prices. This index number is a weighted aggregate of actual prices of 43 foodstuffs. Data are collected from about 50 industrial cities. Foodstuffs included are: sirloin steak, round steak, bacon, eggs, butter, milk, flour, potatoes, sugar, and the like. Weights used are based on an investigation into the amounts of the various articles used by the typical workingman's family. 1913 is used as a base.

**118. Federal Reserve Board.**—The Federal Reserve Board publishes in the *Federal Reserve Bulletin* an index number of wholesale prices. The methods and weights used are the same as those of the Bureau of Labor Statistics. The data used are those compiled by the Bureau of Labor Statistics but regrouped. Index numbers are computed for the following groups:

Consumer's goods
Producer's goods
Raw materials
- Animal products
- Crops
- Forest products
- Mineral products

As an aid in the comparison of international price movements, the Federal Reserve Board is publishing index numbers of wholesale prices for the United States, Canada, France, England, and Japan for the following groups of commodities:

Consumer's goods.
Producer's goods.
Raw materials.
Goods produced.
Goods exported.
Goods imported.

**119. Bradstreet's index number.**—In *Bradstreet's* is published a monthly index of wholesale prices. This index is the sum of the actual prices per pound, on the first of the month, of 96 commodities. There is no base period. There are no weights assigned. Some important commodities are given more than one quotation. This index number is a good barometer of business conditions because of the weights implicitly given to raw materials. It was first published in 1892.

**120. Dun's index number.**—*Dun's Review* publishes monthly, about the middle of the month, an index number of wholesale prices. This index is a weighted aggregate of actual prices on the first of the month. It represents the actual cost in dollars and cents to a single individual of a year's supply of certain commodities. There is no base period. Weights used are the average annual per capita consumption of the various commodities. The exact weights used have never been announced. There has never been published a list of the commodities included, although it is known that some 300 are used. The weight given to food products is around 50 per cent of the total. This index number was first published in 1901.

**121. Fisher's index.**—Prof. Irving Fisher publishes weekly, in the press, on Monday, an index number of wholesale prices for the week ending on the previous Friday at noon. This index number is a weighted aggregate of actual prices, expressed as a relative, with 1913 as a base. The weights used are the quantities marketed in 1919 as given by the Census. The prices used are those of 205 articles as given in *Dun's Review*. This index number was first published in January, 1923.

**122. Persons' index.**—Prof. Warren M. Persons has constructed a "commodity price index of business cycles." This is an unweighted geometric mean of relative prices with 1919 as a base. The commodities included are ten in number: cotton-seed oil, coke, bar iron, pig iron, pig zinc, hides, mess pork, print cloth, sheetings, and worsted yarns. These commodities were selected because their price changes in the past were found to correspond most closely with the changes in general business conditions during the period 1903–1914, as determined by the general agreement over this period of several existing index numbers in placing the crests of the wave-like movements of prices in 1907, 1910, and 1912.

**123. Annalist's index.**—*The Annalist* publishes weekly an index of wholesale food prices, based on the price at wholesale of 25 articles of food making up a theoretical family food-budget. This index is a simple arithmetic average of relative prices, unweighted, with the years 1890–99 as a base. It is limited in value.

### F. Foreign Index Numbers.

**124. Foreign index numbers.**—There are also index numbers of importance issued outside of the United States:

Department of Labor, Ottawa, Canada.

*Economist*, London, England.

*The Statist*, London, England.

*Annuaire Statistique de la France*, Paris.

*Jahrbücher für Nationalökonomie und Statistik*, Jena, Germany.

Those who are interested in the details of foreign index numbers are referred to the Bureau of Labor Statistics bulletin number 284, pp. 175–343.

## Chapter XI

# SEASONAL AND CYCLICAL FLUCTUATIONS

**125. Elements of a fluctuation.**—With statistical data, it frequently happens that one of the variables is time. This is especially true of data relating to economic and sociological phenomena. Such data are sometimes referred to as time series or historical series. Such a series, when plotted, may present a curve which appears to the unaided eye to possess some more or less regular fluctuations, or to follow more or less roughly some direction or trend.

It has been found possible to break up such a series into a number of constituent parts. These parts are known as (a) the secular trend, (b) seasonal variation, (c) long-time cycle.

In addition there are sometimes found *irregular* deviations due to wars, financial panics, new inventions, progress in marketing and transportation, and strikes. To illustrate, the price of corn is subject to seasonal fluctuations. The low price for the year is usually shortly after harvest in the fall, and the high price for the year is usually in the summer shortly before harvest. Investigation also discloses that the price of corn is subject to a long-time cycle of about ten years. For example, the average farm price on December 1, 1896,[1] was 21.5 cents, the lowest price between the high price of 50.6 cents in 1890 and the high price of 60.5 cents in 1901. On December 1, 1906, the price was 39.9 cents, the lowest price between the high price of 60.5 cents in 1901 and the high price of 61.8 cents in 1911. We notice also that, during this period of twenty years, there is a general tendency toward an increase in price, which we call an upward secular trend.

Let us study these movements in turn.

[1] Yearbook of the U. S. Department of Agriculture, 1923.

### Exercise.

The student should plot the prices for the different farm products given in tables XV and XVI, appendix.

## A. Secular Trend.

**126. Graphical determination.**—The plotted data give, in general, a wave-like curve. Stretch a thread along this curve in such a way that the area between the curve and the thread is distributed in approximately equal amounts above and below the thread. A peak and an adjacent trough should be at about the same distance above and below the thread. The straight line thus indicated gives a fair representation of the trend. We obtain in this way the straight line $C$ in fig. 45. (See p. 203.)

**127. Moving average.**—Table 72 gives the monthly average price per dozen of strictly fresh eggs in the United States for the years 1920, 1921, 1922.[2]

**Table 72.—Price per Dozen of Strictly Fresh Eggs, and Moving Averages.**

| *Yr. and Mo.* | *Price ¢* | *3 Mo. Av.* | *12 Mo. Av.* | *Yr. and Mo.* | *Price ¢* | *3 Mo. Av.* | *12 Mo. Av.* |
|---|---|---|---|---|---|---|---|
| *1920* Jan....... | 82.7 | .... | .... | July....... | 42.3 | 41.6 | 50.9 |
| Feb....... | 68.5 | 68.9 | .... | Aug....... | 47.6 | 46.8 | 48.5 |
| March.... | 55.6 | 59.0 | .... | Sept....... | 50.4 | 52.3 | 48.5 |
| April...... | 52.8 | 53.8 | .... | Oct....... | 58.9 | 59.6 | 47.7 |
| May...... | 52.9 | 53.1 | .... | Nov....... | 69.5 | 66.3 | 47.5 |
| June...... | 53.6 | 54.6 | .... | Dec....... | 70.5 | 63.3 | 47.5 |
| July...... | 57.3 | 58.2 | 68.1 | *1922* Jan....... | 49.9 | 56.3 | 47.4 |
| Aug...... | 63.6 | 64.0 | 67.8 | Feb....... | 48.4 | 43.4 | 46.9 |
| Sept...... | 71.1 | 71.8 | 66.1 | March..... | 31.8 | 37.3 | 46.0 |
| Oct....... | 80.8 | 79.3 | 64.9 | April...... | 31.7 | 32.3 | 45.5 |
| Nov...... | 86.1 | 86.4 | 63.4 | May...... | 33.5 | 33.1 | 45.1 |
| Dec....... | 92.4 | 85.9 | 61.8 | June...... | 34.1 | 34.5 | 44.7 |
| *1921* Jan....... | 79.1 | 73.1 | 60.2 | July....... | 36.0 | 35.7 | 44.4 |
| Feb....... | 47.9 | 56.2 | 59.0 | Aug....... | 37.1 | 39.3 | |
| March.... | 41.7 | 41.3 | 57.6 | Sept....... | 44.8 | 45.4 | |
| April...... | 34.3 | 36.5 | 55.9 | Oct....... | 54.3 | 54.5 | |
| May...... | 33.4 | 34.2 | 54.9 | Nov....... | 64.5 | 61.8 | |
| June...... | 35.0 | 36.9 | 52.7 | Dec....... | 66.5 | | |

[2] United States Bureau of Labor Statistics, Bulletin 334, p. 81.

We have computed the moving average for three months, centered on the middle month and tabulated in column three. That is, we add the prices for the first three months, divide by three, and place this average opposite February. Then we average the prices for February, March, and April and tabulate for March. The computation may be shortened by noting the difference between the price dropped and the price added, dividing the difference by the number of months averaged, and adding this quotient, sign considered, to the last computed average.

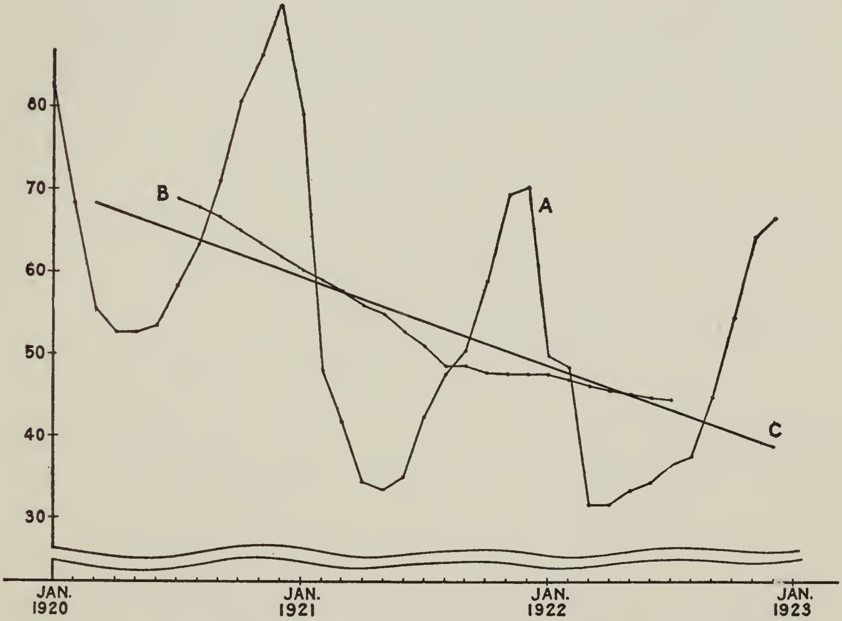


Fig. 45.—Cyclical Fluctuation in Price of Fresh Eggs.

The plotted data give us the wave-like curve *A* in fig. 45. This curve shows a *seasonal variation* with high prices in December and January, followed by a quick drop in prices to a low point in March, April, or May, and then a gradual rise to a high price again in December or January. In addition, there appears to be a general downward movement in price during these three years. This general downward movement we will obtain by the moving average method.

The moving average of three months shows nearly as much seasonal variation as the original data. To eliminate this seasonal variation, we must take a period equal to the typical period for the data under consideration. For the price of eggs, obviously, the period is twelve months. We then compute a moving average for twelve months and center on the seventh month. The results of this computation are shown in column four. The data of column four, when plotted, give the curve $B$ in fig. 45.

**128. Fitting straight line to trend.**—We will illustrate the method by a simple numerical example. Let it be required to fit a straight line to the data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $X$: | 1 | 2 | 4 | 6 | 8 | 9 | 10 |
| $Y$: | 3 | 4 | 5 | 7 | 8 | 9 | 10 |

We assume that

$$Y = mX + k$$

From this equation we find that for

$$\begin{aligned} X &= 1 & Y &= 1 \cdot m + k \\ X &= 2 & Y &= 2 \cdot m + k \\ X &= 4 & Y &= 4 \cdot m + k \\ X &= 6 & Y &= 6 \cdot m + k \\ X &= 8 & Y &= 8 \cdot m + k \\ X &= 9 & Y &= 9 \cdot m + k \\ X &= 10 & Y &= 10 \cdot m + k \end{aligned}$$

Adding, we find that the sum of the computed values of $Y$ is

$$40m + 7k$$

The sum of the measured values of $Y$ is

$$3 + 4 + 5 + 7 + 8 + 9 + 10 = 46$$

Let us agree to determine $m$ and $k$ in such a way that these sums shall be equal. Then

$$40m + 7k = 46 \tag{1}$$

Let us multiply each computed value of $Y$ by the corresponding value of $X$ and add. This is the first moment of the computed value of $Y$. We have

$$\begin{array}{rl} & 1(\ 1m + k) \\ & 2(\ 2m + k) \\ & 4(\ 4m + k) \\ & 6(\ 6m + k) \\ & 8(\ 8m + k) \\ & 9(\ 9m + k) \\ & 10(10m + k) \\ \hline \text{Adding:} & 302m + 40k \end{array}$$

Multiply each measured value of $Y$ by the corresponding value of $X$ and add. This is the *first moment* of the measured values of $Y$. We have

$$1 \cdot 3 + 2 \cdot 4 + 4 \cdot 5 + 6 \cdot 7 + 8 \cdot 8 + 9 \cdot 9 + 10 \cdot 10 = 318$$

Let us agree to determine $m$ and $k$ in such a way that these first moments shall be equal. Then

$$302m + 40k = 318 \tag{2}$$

We have thus two linear equations, (1) and (2), for the determination of $m$ and $k$. Solving, we have

$$Y = 0.751X + 2.28$$

In general, if we desire to fit a straight line

$$Y = mX + k$$

to the points $(X_1, Y_1)$, $(X_2, Y_2)$, . . . , $(X_n, Y_n)$, we arrive at the two following equations for the determination of $m$ and $k$:

$$\begin{aligned} \sum_1^n {}_i Y_i &= \sum_1^n {}_i(mX_i + k) = m\sum_1^n {}_i X_i + nk \\ \sum_1^n {}_i X_iY_i &= \sum_1^n {}_i (mX_i + k)X_i = m \sum_1^n {}_i X_i^2 + k \sum_1^n {}_i X_i \end{aligned} \tag{3}$$

The computations are much simplified if we take as origin the arithmetic mean of the $X_i$. To make this change of origin, put

$$x = X - \overline{X};\ y = Y, \text{ where } \overline{X} = \frac{\Sigma X_i}{n}$$

Then $\Sigma x_i = 0$, and our equations (3) become

$$\Sigma y_i = nk,\ \Sigma x_i y_i = m\Sigma x_i^2,$$

whence

$$k = \frac{\Sigma y_i}{n},\ m = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

and

$$y = \frac{\Sigma x_i y_i}{\Sigma x_i^2} x + \frac{\Sigma y_i}{n} \tag{4}$$

$\Sigma x_i^2$ can be computed most advantageously, whenever $x_i$ takes successive integral values from 1 to $n$, from the formula[3]

$$\sum_1^n x^2 = \frac{n(n+1)(2n+1)}{6}$$

The application of formula (4) to the data for the price of eggs in 1920–22 is given in table 73. Substituting the totals from the tables, we find

$$k = \frac{\Sigma y}{n} = \frac{1894.1}{35} = 54.1; m = \frac{-2,893}{3,570} = -0.81,$$

whence

$$y = -0.81x + 54.1 \tag{5}$$

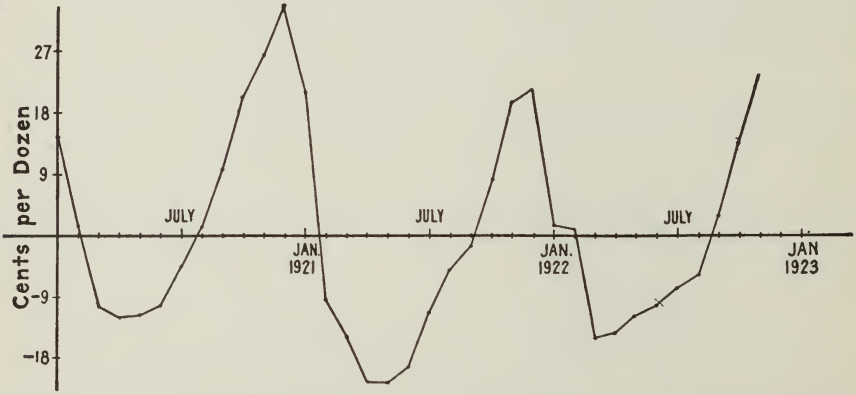


**Fig. 46.—Seasonal Fluctuation in the Price of Eggs, with the Secular Trend Removed. Data, Table 72. Computations, Table 73.**

Having obtained the equation of the straight line (5) which gives the trend, one can, for the given values of $x$, compute the ordinates $Y$ of the secular trend. If one computes the difference $y - Y$ of the original prices of eggs and the prices computed from the secular trend, one obtains the fluctuations in price due to the seasonal and other variations, with the secular trend eliminated. For the prices of eggs, the fluctuations, after the secular trend has been eliminated,

[3] Rietz and Crathorne, "College Algebra," p. 87.

are shown numerically in the last column of table 73, and are shown graphically in fig. 46.

**Table 73.—Determination of the Line of Secular Trend and Deviations from Trend for the Prices of Eggs Given in Table 72.**

| *Date* | $x$ | $y =$ *Price* | $xy$ | $x^2$ | *Ordinate of Secular Trend* $Y$ | *Deviation from Trend* $y - Y$ |
|---|---|---|---|---|---|---|
| *1920* Jan. | −17 | 82.7 | −1,405.9 | 289 | 67.9 | 14.8 |
| Feb. | −16 | 68.5 | −1,096.0 | 256 | 67.1 | 1.4 |
| March | −15 | 55.6 | − 834.0 | 225 | 66.2 | −10.6 |
| April | −14 | 52.8 | − 739.2 | 196 | 65.4 | −12.6 |
| May | −13 | 52.9 | − 687.7 | 169 | 64.6 | −11.7 |
| June | −12 | 53.6 | − 643.2 | 144 | 63.8 | −10.2 |
| July | −11 | 57.3 | − 630.3 | 121 | 63.0 | − 5.7 |
| Aug. | −10 | 63.6 | − 636.0 | 100 | 62.2 | 1.4 |
| Sept. | − 9 | 71.1 | − 639.9 | 81 | 61.4 | 9.7 |
| Oct. | − 8 | 80.8 | − 646.4 | 64 | 60.6 | 20.2 |
| Nov. | − 7 | 86.1 | − 602.7 | 49 | 59.8 | 26.3 |
| Dec. | − 6 | 92.4 | − 554.4 | 36 | 59.0 | 33.4 |
| *1921* Jan. | − 5 | 79.1 | − 395.5 | 25 | 58.2 | 20.9 |
| Feb. | − 4 | 47.9 | − 191.6 | 16 | 57.3 | − 9.4 |
| March | − 3 | 41.7 | − 125.1 | 9 | 56.5 | −14.8 |
| April | − 2 | 34.3 | − 68.6 | 4 | 55.7 | −21.4 |
| May | − 1 | 33.4 | − 33.4 | 1 | 54.9 | −21.5 |
| June | 0 | 35.0 | 000.0 | 0 | 54.1 | −19.1 |
| July | 1 | 42.3 | 42.3 | 1 | 53.3 | −11.0 |
| Aug. | 2 | 47.6 | 95.2 | 4 | 52.5 | − 4.9 |
| Sept. | 3 | 50.4 | 151.2 | 9 | 51.7 | − 1.3 |
| Oct. | 4 | 58.9 | 235.6 | 16 | 50.9 | 8.0 |
| Nov. | 5 | 69.5 | 347.5 | 25 | 50.0 | 19.5 |
| Dec. | 6 | 70.5 | 423.0 | 36 | 49.2 | 21.3 |
| *1922* Jan. | 7 | 49.9 | 349.3 | 49 | 48.4 | 1.5 |
| Feb. | 8 | 48.4 | 387.2 | 64 | 47.6 | 0.8 |
| March | 9 | 31.8 | 286.2 | 81 | 46.8 | −15.0 |
| April | 10 | 31.7 | 317.0 | 100 | 46.0 | −14.3 |
| May | 11 | 33.5 | 368.5 | 121 | 45.2 | −11.7 |
| June | 12 | 34.1 | 409.2 | 144 | 44.4 | −10.3 |
| July | 13 | 36.0 | 468.0 | 169 | 43.6 | − 7.6 |
| Aug. | 14 | 37.1 | 519.4 | 196 | 42.8 | − 5.7 |
| Sept. | 15 | 44.8 | 672.0 | 225 | 42.0 | 2.9 |
| Oct. | 16 | 54.3 | 868.8 | 256 | 41.1 | 13.2 |
| Nov. | 17 | 64.5 | 1,096.5 | 289 | 40.3 | 24.2 |
| TOTAL | 0 | 1,894.1 | −2,893 | 3,570 | | |

**129. Straight line trend by method of least squares.**—The straight line which best fits the given data is generally considered to be that straight line for which the square of the differences of the measured ordinates $y_i$ and the computed ordinates $y_i'$ is a minimum. We have

$$\Sigma(y_i' - y_i)^2 = \Sigma(mx_i + k - y_i)^2$$

$$= m^2\Sigma x_i^2 - 2m\Sigma x_i y_i + (nk^2 - 2k\Sigma y_i + \Sigma y_i^2) \tag{6}$$

$$= nk^2 - 2k\Sigma y_i + (m^2\Sigma x_i^2 - 2m\Sigma x_i y_i + \Sigma y_i^2) \tag{7}$$

if we take as origin for $x$ the arithmetic mean value of $x$, whence $\Sigma x_i = 0$

Equation (6) is a quadratic in $m$ while (7) is a quadratic in $k$. We desire the values of $m$ and $k$ for which these quadratics have a minimum value. Now the general quadratic $ax^2 + bx + c$ can be put in the form

$$a\left(x + \frac{b}{2a}\right)^2 + \frac{4ac - b^2}{4a}$$

This quadratic expression obviously has a minimum if $x = \frac{-b}{2a}\cdot$ Applying this result to the two quadratics, (6) and (7), we find that $\Sigma(y_i' - y_i)^2$ is a minimum when

$$m = \frac{\Sigma x_i y_i}{\Sigma x_i^2}, \quad k = \frac{\Sigma y_i}{n} \tag{8}$$

But these are the same values obtained in the preceding section by the method of moments. Thus we see that the best fitting straight line in the sense of least squares is the straight line obtained by the method of moments.[3]

**130. Method of semi-averages for a straight-line trend.**—The range for $x$ is divided into two parts, as nearly equal as possible. For each part, compute the arithmetic average value of the corresponding ordinates. Plot these mean values of $y$ at the mid-point of their respective $x$-ranges.

[3] Those acquainted with the calculus can derive equations (8) from the following considerations: Equation (6) is a function of the two variables $m$ and $k: f(m, k)$. This function will be a minimum if

$$\frac{\partial f}{\partial m} = 0 \text{ and } \frac{\partial f}{\partial k} = 0.$$

These two partial derivatives give

$$\Sigma(mx_i + k - y_i)\, x_i = 0 \text{ or } \Sigma x_i d_i = 0$$

and

$$\Sigma(mx_i + k - y_i) = 0 \text{ or } \Sigma d_i = \Sigma(y_i' - y_j) = 0.$$

These two partial derivatives, when solved for $m$ and $k$, give equations (8).

Join the points so found by a straight line This straight line will be found to be a fair fit to the moving average and to the line of least squares. For the table of egg prices, table 73, divide the range into seventeen months and eighteen months. The average price for the first seventeen months is 61.99 cents, which should be plotted as an ordinate for the ninth month, that is, for September, 1920. The average price for the last eighteen months is 46.68 cents, which should be plotted midway between February and March of 1922.

**131. Non-linear trends.**—The moving average when plotted may not give a straight line at all. For example, the moving average for egg prices from Jan., 1917, to Jan., 1922, would not give a straight line, but a curve resembling a parabola with axis vertical. For the trend in the years 1917–1919 is upward[4] and in the years 1920–1922 the trend is downward. For this case we should try to fit a curve of the form

$$y = a + bx + cx^2$$

The three linear equations for the determination of $a$, $b$, and $c$ are

$$\begin{aligned}\Sigma y_i &= \Sigma(a + bx_i + cx_i^2)\\ \Sigma x_i y_i &= \Sigma x_i(a + bx_i + cx_i^2)\\ \Sigma x_i^2 y_i &= \Sigma x_i^2(a + bx_i + cx_i^2)\end{aligned}$$

If we take the mean of the $x$ for origin, the equations are simplified somewhat, and become

$$\begin{aligned}\Sigma y_i &= na + c\Sigma x_i^2\\ \Sigma x_i y_i &= \qquad b\Sigma x_i^2\\ \Sigma x_i^2 y_i &= \qquad a\Sigma x_i^2 + c\Sigma x_i^4,\end{aligned}$$

since

$$\Sigma x_i = 0 \text{ and } \Sigma x_i^3 = 0$$

The extension to the case where one tries to fit a curve of the form

$$y = a_o + a_1x + a_2x^2 + \cdots + a_nx^n$$

would seem to be obvious.

In general, it does not seem to be desirable or advantageous to fit a trend with a curve of degree higher than the second.

[4] Jerome, Harry, "Statistical Method," p. 228. Harpers', 1924.

Tables[5] are given for $\Sigma x$, $\Sigma x^2$, $\Sigma x^3$, $\Sigma x^4$.

In various college algebras,[6] one can find the following formulas for the computation of these summations:

$$\sum_{1}^{n}{}_{x}\, x = \frac{n(n+1)}{2}$$

$$\sum_{1}^{n}{}_{x}\, x^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{1}^{n}{}_{x}\, x^3 = \frac{n^2(n+1)^2}{4}$$

$$\sum_{1}^{n}{}_{x}\, x^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}$$

To illustrate, let us fit a curve of the type

$$y = a + bx + cx^2$$

to the data

| $x$: | $-2$ | $-1$ | 0 | 1 | 2 |
|---|---|---|---|---|---|
| $y$: | 7 | 11 | 13 | 10 | 8 |

We have

$$\Sigma y = 49;\ \Sigma xy = 1;\ \ \Sigma x^2y = 81$$
$$\Sigma x = 0;\ \ \Sigma x^2 = 10;\ \ \Sigma x^3 = 0;\ \Sigma x^4 = 34$$

Whence, for the determination of $a$, $b$, and $c$, we have

$$5a + 10c = 49$$
$$10a + 34c = 81$$
$$10b = 1$$

Whence

$$y = 12.2 + 0.1x - 1.2x^2$$

**132. Selected points.**—Determine whether a straight line, a parabola, or some other form of curve should be chosen to represent the trend. Draw roughly, free hand, an experimental curve among the plotted points. Select as many points, near this experimental curve, as there are constants to be determined. Thus, for a straight line select two, for a parabola three, and so on. Determine the constants in

[5] Table XXVIII, Pearson, "Tables for Statisticians, and Biometricians," Cambridge University Press, 1896.

[6] For example, Fine, "College Algebra," p. 370, Ginn and Co.

such a way that the required line of trend passes exactly through the selected points.

### Exercises.

1. From the following data plot the seasonal fluctuations of eggs (*Source: U. S. Bureau of Labor Statistics, Bulletin 270, pp. 62–65*):

Average price per dozen of strictly fresh eggs in the U. S.

| | Jan. | Feb. | March | April | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *1917* | 55 | 51 | 35 | 39 | 40 | 41 | 42 | 46 | 53 | 55 | 58 | 64 |
| *1918* | 67 | 63 | 44 | 43 | 42 | 43 | 49 | 54 | 59 | 64 | 74 | 81 |
| *1919* | 75 | 51 | 48 | 49 | 53 | 54 | 57 | 60 | 63 | 72 | 81 | 90 |

2. Fit a straight line to the data in exercise 1. Use July 1, 1918 as origin. *Ans.* $y = 0.8x + 56$.

3. Fit a straight line to the data in exercise 1, using the method of semi-averages.

## B. Seasonal Variation.

**133. Link relative.**—In article 127 we observed a seasonal variation in the price of eggs. A certain downward secular trend, for a period of three years, was observed. We obtained, fig. 46, a graph showing the seasonal fluctuations with the secular trend removed. A study of this graph shows that the fluctuations from year to year were not uniform. We desire a *typical* seasonal movement with the trend eliminated. A device known as the *link-relative* method has been constructed for this purpose. After the typical seasonal variation has been found, both it and the secular trend are liminated, leaving the data in suitable shape for a careful study of the fluctuations due to the business cycle.

Table 74 gives the deaths from typhoid fever in New York State by months from January, 1910, to December, 1921, inclusive, as taken from the United States Census Bureau Mortality Statistics. Let us compute from these data the ratio which the deaths for any one month bear to the deaths for the preceding month. This ratio is called the *link relative.* Thus, the ratio of February to January deaths for 1910 is $\frac{92}{91} = 1.01$. The ratio of March deaths to February

Table 74.—Deaths from Typhoid Fever in New York State by Months.

| | Jan. | Feb. | March | April | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. | Totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1910 | 91 | 92 | 97 | 74 | 66 | 72 | 100 | 142 | 185 | 174 | 168 | 136 | 1,397 |
| 11 | 97 | 90 | 94 | 80 | 78 | 75 | 116 | 167 | 154 | 148 | 117 | 104 | 1,320 |
| 12 | 109 | 64 | 69 | 71 | 83 | 73 | 77 | 121 | 138 | 143 | 86 | 73 | 1,107 |
| 13 | 53 | 46 | 52 | 45 | 64 | 48 | 72 | 105 | 126 | 168 | 121 | 99 | 999 |
| 1914 | 70 | 56 | 53 | 62 | 63 | 45 | 70 | 74 | 114 | 117 | 94 | 66 | 884 |
| 15 | 51 | 37 | 45 | 36 | 43 | 50 | 52 | 89 | 97 | 101 | 77 | 87 | 765 |
| 16 | 60 | 39 | 33 | 33 | 33 | 39 | 41 | 61 | 72 | 82 | 67 | 51 | 611 |
| 17 | 41 | 47 | 39 | 37 | 40 | 34 | 46 | 65 | 86 | 63 | 55 | 36 | 589 |
| 1918 | 28 | 21 | 34 | 29 | 42 | 31 | 53 | 58 | 75 | 140 | 35 | 33 | 579 |
| 19 | 24 | 21 | 22 | 16 | 26 | 14 | 34 | 44 | 58 | 49 | 30 | 36 | 374 |
| 20 | 16 | 12 | 13 | 30 | 17 | 23 | 43 | 49 | 68 | 48 | 27 | 33 | 379 |
| 21 | 28 | 15 | 27 | 17 | 16 | 26 | 34 | 42 | 52 | 60 | 34 | 35 | 386 |

*U. S. Census Bureau, Mortality Statistics.*

deaths is $\frac{97}{92}$ = 1.05. That is, the link relative for February is 1.01. The link relative for March is 1.05. The link relatives for the years 1910–1920, inclusive, are given in table 75.

What we desire is a measure of the typical seasonal fluctuation. The link relatives for January vary from year to year. Let us take as the typical link relative for January an average of the January link relatives. The median seems to be the best average to use. This modifies the influence of extreme items, which are often due to exceptional circumstances. The median for each month is given in the

Table 75.—Link Relatives for Deaths from Typhoid Fever in New York State.

| | Jan. | Feb. | March | April | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1910 | .... | 1.01 | 1.05 | .76 | .89 | 1.09 | 1.39 | 1.42 | 1.30 | .94 | .97 | .81 |
| 11 | .71 | .93 | 1.04 | .85 | .97 | .96 | 1.55 | 1.44 | .92 | .96 | .79 | .89 |
| 12 | 1.05 | .59 | 1.08 | 1.03 | 1.17 | .88 | 1.05 | 1.57 | 1.14 | 1.04 | .60 | .85 |
| 13 | .73 | .87 | 1.13 | .87 | 1.42 | .75 | 1.50 | 1.46 | 1.20 | 1.33 | .72 | .82 |
| 1914 | .71 | .80 | .95 | 1.17 | 1.02 | .71 | 1.55 | 1.06 | 1.54 | 1.03 | .80 | .70 |
| 15 | .77 | .73 | 1.21 | .80 | 1.19 | 1.16 | 1.04 | 1.71 | 1.09 | 1.04 | .76 | 1.13 |
| 16 | .69 | .65 | .85 | 1.00 | 1.00 | 1.18 | 1.05 | 1.49 | 1.18 | 1.14 | .82 | .76 |
| 17 | .80 | 1.15 | .83 | .95 | 1.08 | .85 | 1.35 | 1.41 | 1.32 | .73 | .87 | .65 |
| 1918 | .78 | .75 | 1.62 | .85 | 1.45 | .74 | 1.71 | 1.09 | 1.29 | 1.87 | .25 | .94 |
| 19 | .73 | .88 | 1.05 | .73 | 1.62 | .54 | 2.42 | 1.29 | 1.32 | .84 | .61 | 1.20 |
| 20 | .44 | .78 | 1.08 | 2.31 | .57 | 1.35 | 1.87 | 1.14 | 1.39 | .71 | .56 | 1.22 |
| Median.. | .73 | .80 | 1.05 | .87 | 1.08 | .88 | 1.50 | 1.42 | 1.29 | 1.03 | .76 | .85 |

last row of table 75. These numbers are called the *median link relatives.* Let us now reduce these percentages to January as a base. The number for January is then 1.000. Since the February deaths are 0.80 of the January number, the number for February is 0.800. Since the March deaths are 1.05 of the February deaths, the March deaths are 0.800 × 1.05 = 0.840 of the January deaths. These percentages of the January deaths are termed *crude serial relatives.* The crude serial relative for any month is obtained as the product of the median link relative for that month and the crude serial relative for the preceding month.

**Table 76.—Computation of Seasonal Base Number from Median Link Relatives. Typhoid Fever in New York State. (See Tables 74 and 75.)**

| *A* | *B* | *C* | *D* | *E* | *F* | *G* |
|---|---|---|---|---|---|---|
| | *Median Link Relative* | *Crude Serial Relative* | *Monthly Adjustment for Secular Trend* .006 | *Adjusted Serial Relative* | *Serial Relative Average for Year as Base* | *Seasonal Base Number Average for 1920 as Base* |
| Jan. | .73 | 1.000 | .... | 1.000 | .833 | 26 |
| Feb. | .80 | .800 | .006 | .806 | .671 | 21 |
| March | 1.05 | .840 | .012 | .852 | .709 | 22 |
| April | .87 | .731 | .018 | .749 | .624 | 20 |
| May | 1.08 | .789 | .025 | .814 | .678 | 21 |
| June | .88 | .695 | .031 | .726 | .604 | 19 |
| July | 1.50 | 1.042 | .037 | 1.079 | .898 | 28 |
| Aug. | 1.42 | 1.479 | .043 | 1.522 | 1.267 | 40 |
| Sept. | 1.29 | 1.908 | .049 | 1.957 | 1.629 | 51 |
| Oct. | 1.03 | 1.965 | .055 | 2.020 | 1.682 | 53 |
| Nov. | .76 | 1.493 | .062 | 1.555 | 1.295 | 41 |
| Dec. | .85 | 1.269 | .068 | 1.337 | 1.113 | 35 |
| Jan. | .73 | .926 | .074 | 1.000 | | |

If we multiply the December crude serial relative by the January median link relative, we obtain in this case 0.926, which does not agree with the number 1.000 with which we started. This difference is due to the secular trend. If the secular trend is linear, this difference is cumulative and additive. That is, the difference between January and December is twelve times the difference between January and February. Then the difference $1.000 - 0.926 = 0.074$ should be added, $\frac{1}{12}$ to February, $\frac{2}{12}$ to March, $\frac{3}{12}$ to April, and so on.

This adjustment will make both January serial relatives read 1.000. The resulting numbers given in column *E* of table 76 are the serial relatives adjusted for secular trend. These numbers constitute a typical seasonal death rate expressed with January as a base.

Now January might be above or below the average for the year. It seems desirable to express the deaths with the average for the year as a base. The arithmetic average for the year is 1.2014. We compute the ratio of each number in column *E* to 1.2014 and place the result in column *F*. These numbers are typical ratios of the deaths per month with the average for the year as a base.

The average number of actual deaths per month for 1920 is 31.6. Multiplying this number by the numbers in column *F* gives the numbers in column *G*, which are called *seasonal base numbers*. These numbers agree closely with the actual number of deaths tabulated in table 74 for either the year 1919 or 1921.

### Exercise.

1. Compute seasonal base numbers for the data in (a) table XVI, appendix; and (b) table 72, and the table in exercise 1, article 132.

## C. Long Time Cycles.

**134. Business cycles.**—Prices, wages, production in both industrial and agricultural lines, and many other series connected with business activities, are affected by the alternating periods of depression and prosperity which occur in business. The lengths of these periods may and do vary, but there is sufficient regularity present to enable one to study these movements as cyclical phenomena. From the viewpoint of the business man the study of these movements is valuable inasmuch as they are indicators of business conditions.

Data for the study of these movements may be given by months over a period of years or there may be given but one number for each year. For example, the *Yearbook of the Department of Agriculture* gives the average farm price of corn on Dec. 1 from 1870 down to the present and also the yearly production in millions of bushels. In the *Statistical Abstract of the United States* are given annual data on production of pig iron and bituminous coal in millions of tons, New York bank clearings in millions of dollars, liabilities in millions of dollars of commercial failures, and other data on production. The *Bureau of Labor Statistics Bulletin No. 334* gives the average retail prices by months for a number of years for a great variety of articles such as eggs, butter, pork chops, bituminous coal, cotton cloth of various grades, building materials, and so on. *Bulletin 335* of the same bureau gives the average wholesale prices by months for the same articles.

In the past, major periods of depression occurred in 1819, 1837, 1857, 1873, 1893, 1907. There have been a number of notable attempts[7] to connect these periods of depression with the periods of maximum and minimum rainfall and these periods of rainfall with the periods of maximum and minimum number of sun spots.

**135. Secular trend.**—Whether the data are given by months or years, the computation of the trend, whether straight line or otherwise, is the same as outlined under secular trend for seasonal variations. The secular trend can be eliminated in the manner previously outlined for seasonal data.

Table 77A gives the average farm price on Dec. 1 of corn and wheat. The table contains all necessary computations for the straight line trend. The equations for the trend are below the table.

[7] See, for example, Moore, H. L., "Economic Cycles: Their Law and Cause." Macmillan, New York, 1914.

*Monthly Weather Review*, March 1922, pp. 128–130. Rainfall departures in inches from the mean, for the U. S. as a unit, and for Arizona, and the number of sunspots for the years indicated.

**Table 77A.—Computation of Secular Trend for Prices of Corn and Wheat.**

$Y = 1.96x + 55.49$ $\qquad\qquad$ $Y = 3.00x + 92.66$

| | | *Corn* | | | | *Wheat* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Yr. | $x$ | Average Farm Price Dec. 1 $y$ | $xy$ | Ordinate of Trend $Y$ | Deviation from Trend $y - Y$ | Average Farm Price Dec. 1 $y$ | $xy$ | $Y$ | $y - Y$ | $x^2$ |
| 1890 | −16 | 50.6 | −809.6 | 24.1 | 26.5 | 83.8 | −1,340.8 | 43.66 | 40.1 | 256 |
| 91 | −15 | 40.6 | −609.0 | 26.1 | 14.5 | 83.9 | −1,258.5 | 47.66 | 36.2 | 225 |
| 92 | −14 | 39.4 | −551.6 | 28.1 | 11.3 | 62.4 | − 873.6 | 50.66 | 11.7 | 196 |
| 93 | −13 | 36.5 | −474.5 | 30.0 | 6.5 | 53.8 | − 699.4 | 53.66 | 0.1 | 169 |
| 94 | −12 | 45.7 | −548.4 | 32.0 | 13.7 | 49.1 | − 589.2 | 56.66 | − 7.6 | 144 |
| 1895 | −11 | 25.3 | −278.3 | 33.9 | − 8.6 | 50.9 | − 559.9 | 59.66 | − 8.8 | 121 |
| 96 | −10 | 21.5 | −215.0 | 35.9 | −14.4 | 72.6 | − 726.0 | 62.66 | + 9.9 | 100 |
| 97 | − 9 | 26.3 | −236.7 | 37.9 | −11.6 | 80.8 | − 727.2 | 65.66 | 15.1 | 81 |
| 98 | − 8 | 28.7 | −229.6 | 39.8 | −11.1 | 58.2 | − 465.6 | 68.66 | −10.5 | 64 |
| 99 | − 7 | 30.3 | −212.1 | 41.8 | −11.5 | 58.4 | − 408.8 | 71.66 | −13.3 | 49 |
| 1900 | − 6 | 35.7 | −214.2 | 43.7 | − 8.0 | 62.0 | − 372.0 | 74.66 | −12.7 | 36 |
| 01 | − 5 | 60.5 | −302.5 | 45.7 | +14.8 | 62.6 | − 313.0 | 77.66 | −15.1 | 25 |
| 02 | − 4 | 40.3 | −161.2 | 47.7 | − 7.4 | 63.0 | − 252.0 | 80.66 | −17.7 | 16 |
| 03 | − 3 | 42.5 | −127.5 | 49.6 | − 7.1 | 69.5 | − 208.5 | 83.66 | −14.2 | 9 |
| 04 | − 2 | 44.1 | − 88.2 | 51.6 | − 7.5 | 92.4 | − 184.8 | 86.66 | + 5.7 | 4 |
| 1905 | − 1 | 41.2 | − 41.2 | 53.5 | −12.3 | 74.6 | − 74.6 | 89.66 | −15.1 | 1 |
| 06 | 0 | 39.9 | 0 | 55.5 | −15.6 | 66.2 | 0 | 92.66 | −26.5 | 0 |
| 07 | 1 | 51.6 | 51.6 | 57.5 | − 5.9 | 86.5 | + 86.5 | 95.66 | − 9.2 | 1 |
| 08 | 2 | 60.6 | 121.2 | 59.4 | + 0.8 | 92.2 | 184.4 | 98.66 | − 6.5 | 4 |
| 09 | 3 | 57.9 | 173.7 | 61.4 | − 3.5 | 98.4 | 295.2 | 101.66 | − 3.3 | 9 |
| 1910 | 4 | 48.0 | 192.0 | 63.3 | −15.3 | 88.3 | 353.2 | 104.66 | −16.4 | 16 |
| 11 | 5 | 61.8 | 309.0 | 65.3 | − 3.5 | 87.4 | 437.0 | 107.66 | −20.3 | 25 |
| 12 | 6 | 48.7 | 292.2 | 67.3 | −18.6 | 76.0 | 456.0 | 110.66 | −34.7 | 36 |
| 13 | 7 | 69.1 | 483.7 | 69.2 | − 0.1 | 76.9 | 538.3 | 113.66 | −36.8 | 49 |
| 14 | 8 | 64.4 | 515.2 | 71.2 | − 6.8 | 98.6 | 788.8 | 116.66 | −18.1 | 64 |
| 1915 | 9 | 57.5 | 517.5 | 73.1 | −15.6 | 91.9 | 827.1 | 119.66 | −27.8 | 81 |
| 16 | 10 | 88.9 | 889.0 | 75.1 | 13.8 | 160.3 | 1,603.0 | 122.66 | +37.6 | 100 |
| 17 | 11 | 127.9 | 1,406.9 | 77.1 | 50.8 | 200.8 | 2,208.8 | 125.66 | +75.1 | 121 |
| 18 | 12 | 136.5 | 1,638.0 | 79.0 | 57.5 | 204.2 | 2,450.4 | 128.66 | 75.5 | 144 |
| 19 | 13 | 134.5 | 1,748.5 | 81.0 | 53.5 | 214.9 | 2,793.7 | 131.66 | 83.2 | 169 |
| 1920 | 14 | 67.0 | 938.0 | 82.9 | −15.9 | 143.7 | 2,011.8 | 134.66 | + 9.0 | 196 |
| 21 | 15 | 42.3 | 634.5 | 84.9 | −42.6 | 92.6 | 1,389.0 | 137.66 | −45.1 | 225 |
| 22 | 16 | 65.7 | 1,061.2 | 86.9 | −21.2 | 100.9 | 1,614.4 | 140.66 | −39.8 | 256 |
| TOTALS........ | 0 | 1,831.5 | 5,872.6 | .... | ...... | 3,057.8 | 8,983.7 | ...... | ...... | 2,992 |

*Yearbook of Department of Agriculture.*

Expressed as deviations from the trend, the fluctuations in the price of wheat seem to be greater than for corn. This is because the average price of a bushel of wheat is more than that of corn. Our data need to be reduced to a comparable basis. Considerable help is obtained in making the comparisons by expressing the deviations from the trend as percentages of the trend for each series. The usefulness of this method depends upon the simplicity and ease of the computations. Applying this to the data, table 77A, for corn for the year 1890, we find the percentage deviation from the trend to be

$$100(26.5 \div 50.6) = 52.4$$

We give in table 77B the percentage deviation from the trend for a few years.

**Table 77B.—Percentage Deviations from Trend for Corn and Wheat. Computations Based on Data in Table 77A.**

| | Percentage Deviations from Trend | | | | | |
|---|---|---|---|---|---|---|
| | *1890* | *1891* | *1892* | *1893* | *1894* | *1895* |
| *Corn* | 52.4 | 35.7 | 28.6 | 17.8 | 30.0 | −34.0 |
| *Wheat* | 47.9 | 43.1 | 18.7 | 1.9 | −15.5 | −17.3 |

**Exercises.**

1. Plot the lines of trend and the average farm price on Dec. 1 for both corn and wheat. Use data table 77A.

2. For both corn and wheat, plot the deviation from trend. Use data table 77A.

3. For both corn and wheat, complete table 77B up to the year 1923. Use data table 77A.

4. For both corn and wheat, for the years 1890–1922, plot the percentage deviations from trend.

This reduction, however, does not take account of the distribution of the items between maximum and minimum values. The most satisfactory method of reduction to a comparable basis seems to be to express deviations from the trend in multiples of the standard deviation. That is, each deviation from the trend is divided by the standard deviation and the resulting series tabulated and plotted.

A systematic procedure for a reduction to a comparable basis has been given by Prof. Warren M. Persons of Harvard

University in his "Indices of General Business Conditions," published as preliminary volume I in *The Review of Economic Statistics,* 1919. No study of Business Cycles is complete without a study of this book. An outline of the processes to which individual series were subjected is given on p. 37 of this publication and is reproduced here:

> a. The data necessary for the application of the method are homogeneous monthly series of statistics covering a period of, say, fifteen years or more.
>
> b. The linear secular trend is found by fitting a straight line to annual items by the method of least squares. The compound interest curve, parabola, or other curves are used in case a straight line does not give satisfactory results.
>
> c. Indices of seasonal variation are found by taking the medians of month-to-month percentages for each of the twelve months. The medians thus found are progressively multiplied to form a continuous series, the discrepancy due to secular trend is distributed, and the average is made equal to 100.
>
> d. The original items are corrected for secular trend and seasonal variation, as follows: each monthly ordinate of secular trend is multiplied by the index of seasonal variation for that month; the resulting product is subtracted from the corresponding original item, and then expressed as a percentage of the ordinate of secular trend.
>
> e. The percentage deviations of the various series are expressed in terms of their respective standard deviations, in order to secure comparable cyclical fluctuations.

This method is applied by Prof. Persons to fourteen different series for which consecutive monthly data could be found, covering the period 1903–1918. We reproduce here, in table 78, his data for the production of pig iron and for Bradstreet's prices.

The equation of the line of secular trend is obtained with the abscissa $x$ measured in years from the middle of 1903. For pig iron the ordinate $y$ is measured in units of 1,000 gross tons of pig iron produced per month. For Bradstreet's prices the ordinate $y$ is measured in units of one dollar.

$$\text{Line of secular trend[8]} \begin{cases} \text{Pig iron:} & y = 95x + 1{,}461 \\ \text{Bradstreet's prices:} & y = 0.11x + 7.94 \end{cases}$$

[8] Preliminary volume I of *The Review of Economic Statistics*, p. 39.

**Table 78.—Monthly Tonnage of Pig Iron Produced in the United States. Original Items in Units of 1,000 Gross Tons. Secular Trend y = 95x × 1,461. Bradstreet's Monthly Index ot Commodity Prices. Original Items in Units of One Dollar. Secular Trend y = 0.11x + 7.94.**

| *Year and Month* | Pig Iron | | | | Bradstreet's | | | |
|---|---|---|---|---|---|---|---|---|
| | *Production Y* | *Ordinate of Trend y* | *Seasonal Index** | *Per Cent† Deviation: u* | *Production Y* | *Ordinate of Trend y* | *Seasonal Index* | *Per Cent Deviation: v* |
| *1907* | | | | | | | | |
| Jan. | 2,205 | 1,793 | 99 | +24 | 8.92 | 8.31 | 101.3 | +7.0 |
| Feb. | 2,045 | 1,801 | 94 | +19 | 9.00 | 8.32 | 101.0 | +7.7 |
| March | 2,226 | 1,809 | 106 | +17 | 9.13 | 8.33 | 100.5 | +9.1 |
| April | 2,216 | 1,817 | 103 | +19 | 8.96 | 8.34 | 100.4 | +7.1 |
| May | 2,295 | 1,825 | 104 | +21 | 8.94 | 8.35 | 99.8 | +6.6 |
| June | 2,234 | 1,833 | 98 | +24 | 8.99 | 8.36 | 98.7 | +7.1 |
| July | 2,255 | 1,841 | 97 | +26 | 9.04 | 8.37 | 98.5 | +7.6 |
| Aug. | 2,250 | 1,849 | 98 | +23 | 8.93 | 8.38 | 98.8 | +6.2 |
| Sept. | 2,183 | 1,857 | 98 | +19 | 8.83 | 8.39 | 99.5 | +5.0 |
| Oct. | 2,336 | 1,865 | 105 | +20 | 8.85 | 8.40 | 100.0 | +5.1 |
| Nov. | 1,828 | 1,873 | 99 | − 2 | 8.75 | 8.41 | 100.4 | +3.7 |
| Dec. | 1,234 | 1,881 | 100 | −34 | 8.52 | 8.42 | 101.2 | +1.0 |
| *1908* | | | | | | | | |
| Jan. | 1,045 | 1,888 | 99 | −44 | 8.29 | 8.42 | 101.3 | −1.9 |
| Feb. | 1,077 | 1,896 | 94 | −37 | 8.13 | 8.43 | 101.0 | −4.0 |
| March | 1,228 | 1,904 | 106 | −42 | 7.99 | 8.44 | 100.5 | −5.7 |
| April | 1,149 | 1,912 | 103 | −43 | 8.07 | 8.45 | 100.4 | −4.9 |
| May | 1,165 | 1,920 | 104 | −44 | 7.96 | 8.46 | 99.8 | −6.2 |
| June | 1,092 | 1,928 | 98 | −41 | 7.72 | 8.47 | 98.7 | −9.1 |
| July | 1,218 | 1,936 | 97 | −34 | 7.82 | 8.48 | 98.5 | −8.0 |
| Aug. | 1,348 | 1,944 | 98 | −29 | 7.93 | 8.49 | 98.8 | −6.9 |
| Sept. | 1,418 | 1,952 | 98 | −26 | 7.91 | 8.50 | 99.5 | −7.3 |
| Oct. | 1,563 | 1,960 | 105 | −25 | 8.01 | 8.51 | 100.0 | −6.1 |
| Nov. | 1,577 | 1,968 | 99 | −19 | 8.07 | 8.52 | 100.4 | −5.6 |
| Dec. | 1,740 | 1,976 | 100 | −12 | 8.21 | 8.53 | 101.2 | −4.0 |
| *1909* | | | | | | | | |
| Jan. | 1,801 | 1,983 | 99 | − 8 | 8.26 | 8.54 | 101.3 | −3.5 |
| Feb. | 1,703 | 1,991 | 94 | − 8 | 8.30 | 8.55 | 101.0 | −3.1 |
| March | 1,832 | 1,999 | 106 | −14 | 8.22 | 8.56 | 100.5 | −4.3 |
| April | 1,738 | 2,007 | 103 | −16 | 8.32 | 8.57 | 100.4 | −3.1 |
| May | 1,880 | 2,015 | 104 | −11 | 8.30 | 8.58 | 99.8 | −3.5 |
| June | 1,929 | 2,023 | 98 | − 3 | 8.40 | 8.59 | 98.7 | −2.5 |
| July | 2,101 | 2,031 | 97 | + 7 | 8.46 | 8.60 | 98.5 | −1.9 |
| Aug. | 2,246 | 2,039 | 98 | +12 | 8.50 | 8.61 | 98.8 | −1.4 |
| Sept. | 2,385 | 2,047 | 98 | +18 | 8.59 | 8.62 | 99.5 | −0.6 |
| Oct. | 2,600 | 2,055 | 105 | +22 | 8.75 | 8.63 | 100.0 | +1.0 |
| Nov. | 2,547 | 2,063 | 99 | +24 | 8.96 | 8.64 | 100.4 | +3.6 |
| Dec. | 2,635 | 2,071 | 100 | +27 | 9.13 | 8.65 | 101.2 | +5.2 |

* Adjusted monthly indices of seasonal variation are given on pp. 51–53 in *Indices of General Business Conditions.*

† Percentage deviation of original items (*Y*) from secular trend (*y*) corrected for seasonal variation.

Table 79.—Monthly Pig Iron Production, and Bradstreet's Prices Expressed in Units of the Standard Deviation for 1907, 1908, 1909. For Pig Iron, Standard Deviation $\sigma_u = 19.15$. For Bradstreet's Prices Standard Deviation $\sigma_v = 3.68$. Calculation of Correlation Coefficient $r$.

| Year and Month | Percentage Deviation from Trend | | Percentage Deviation Divided by Standard Deviation | |
|---|---|---|---|---|
| | *Pig Iron Production* $u$ | *Bradstreet's Prices* $v$ | *Pig Iron* $\frac{u}{\sigma_u}$ | *Bradstreet's* $\frac{v}{\sigma_v}$ |
| *1907* | | | | |
| Jan. | +24 | | | |
| Feb. | +19 | +7.0 | +1.3 | +1.9 |
| March | +17 | +7.7 | +1.0 | +2.1 |
| April | +19 | +9.1 | +0.9 | +2.5 |
| May | +21 | +7.1 | +1.0 | +1.9 |
| June | +24 | +6.6 | +1.1 | +1.8 |
| July | +26 | +7.1 | +1.3 | +1.9 |
| Aug. | +23 | +6.2 | +1.2 | +1.7 |
| Sept. | +19 | +5.0 | +1.0 | +1.4 |
| Oct. | +20 | +5.1 | +1.1 | +1.4 |
| Nov. | − 2 | +3.7 | −0.1 | +1.0 |
| Dec. | −34 | +1.0 | −1.8 | +0.3 |
| *1908* | | | | |
| Jan. | −44 | −1.9 | −2.3 | −0.5 |
| Feb. | −37 | −4.0 | −2.0 | −1.1 |
| March | −42 | −5.7 | −2.2 | −1.6 |
| April | −43 | −4.9 | −2.2 | −1.3 |
| May | −44 | −6.2 | −2.3 | −1.7 |
| June | −41 | −9.1 | −2.1 | −2.5 |
| July | −34 | −8.0 | −1.8 | −2.1 |
| Aug. | −29 | −6.9 | −1.5 | −1.9 |
| Sept. | −26 | −7.3 | −1.4 | −2.0 |
| Oct. | −25 | −6.1 | −1.3 | −1.7 |
| Nov. | −19 | −5.6 | −1.0 | −1.5 |
| Dec. | −12 | −4.0 | −0.6 | −1.1 |
| *1909* | | | | |
| Jan. | − 8 | −3.5 | −0.4 | −1.0 |
| Feb. | − 8 | −3.1 | −0.4 | −0.8 |
| March | −14 | −4.3 | −0.8 | −1.2 |
| April | −16 | −3.1 | −0.8 | −0.8 |
| May | −11 | −3.5 | −0.6 | −1.0 |
| June | − 3 | −2.5 | −0.2 | −0.7 |
| July | + 7 | −1.9 | +0.4 | −0.5 |
| Aug. | +12 | −1.4 | +0.6 | −0.4 |
| Sept. | +18 | −0.6 | +0.9 | −0.2 |
| Oct. | +22 | +1.0 | +1.1 | +0.3 |
| Nov. | +24 | +3.6 | +1.3 | +1.0 |
| Dec. | +27 | +5.2 | +1.4 | +1.4 |

*Review of Economic Statistics, Prel. Vol. 1, p. 192.*

The line of secular trend for pig iron is for 1903–1916, covering 168 months. The line of trend for Bradstreet's prices is for 1903–1914, covering 144 months.

We illustrate the computations for finding the numbers in the columns headed *Per cent Deviation* by using the data for pig iron for Jan., 1907:

$$\frac{2{,}205 - (1{,}793 \times 0.99)}{1{,}793} = 0.24 \qquad \text{or 24 per cent.}$$

In order to place these series on a strictly comparable basis, the percentage deviations from the secular trend must be divided by the standard deviation. For pig iron, the standard deviation is $\sigma_u = 19.15$. For Bradstreet's prices the standard deviation is $\sigma_v = 3.68$. Thus for pig iron for Jan., 1907, we have

$$24.0 \div 19.15 = 1.3$$

These computations are placed in table 79.

The correlation coefficient for paired items in a time series is given by the formula

$$r = \frac{\Sigma\left(\frac{u}{\sigma_u} \cdot \frac{v}{\sigma_v}\right)}{N}$$

In order to compute this correlation coefficient, one must have at hand the complete range of data for both sets of data. One cannot compute the correlation coefficient from the sample of 36 items in table 79 by a direct application of the formula to the data given.

## D. Lead and Lag.

**136. Lead and lag.**—Let us consider the data in table 80 (p. 222.) Either from the plotted data or from the data themselves one sees that, in general, an increase (decrease) in bank clearings is followed a year later by an increase (decrease) in interest rates on sixty to ninety day commercial paper. In this case the time series for interest rates is said to *lag* behind the time series for bank clearings. Sometimes it is said that the series for bank clearings *leads* that for interest rates.

Table 80.—Bank Clearings: Interest Rates.

| YEAR | *Bank Clearings New York City* $10,000,000,000 | *Average Monthly Interest Rate on* 60–90-*Day Commercial Paper* | YEAR | *Bank Clearings* | *Interest Rate* |
|---|---|---|---|---|---|
| 1903 | 6.6 | 5.5 | 1914 | 8.3 | 4.8 |
| 04 | 6.9 | 4.2 | 15 | 11.1 | 3.5 |
| 05 | 9.4 | 4.4 | 16 | 16.0 | 3.4 |
| 06 | 10.5 | 5.7 | 17 | 17.7 | 4.7 |
| 07 | 8.7 | 6.4 | 18 | 17.9 | 5.9 |
| 08 | 7.9 | 4.4 | 19 | 23.6 | 5.4 |
| 09 | 10.4 | 4.0 | 20 | 24.3 | 7.4 |
| 10 | 9.7 | 5.0 | 21 | 19.4 | 6.5 |
| 11 | 9.2 | 4.0 | 22 | 21.8 | 4.4 |
| 12 | 10.1 | 4.7 | 23 | 21.4 | 5.0 |
| 13 | 9.5 | 5.6 | 24 | 25.0 | 3.9 |

*Bank Clearings, Source: Commercial and Financial Chronicle: Interest Rates: Federal Reserve Board Service.*

The amount of lag can be determined roughly by plotting the two series and then observing the amount by which one curve must be moved in a horizontal direction in order that the peaks and troughs of the two series shall approximately coincide throughout.

The standard mathematical device, at the present time, in the absence of any better method, for computing the amount of lag, is to compute the correlation coefficients between the two time series for different amounts of lag. Professor Persons has computed coefficients of correlation for pig-iron production and interest rates on sixty to ninety day commercial paper for varying amounts of lag of interest rates behind pig-iron production. The following results were obtained:[9]

| *Lag of Interest Rates in Months* | *0* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *12* |
|---|---|---|---|---|---|---|---|---|---|
| *Coefficient of Correlation*........ | 0.34 | 0.67 | 0.72 | 0.75 | 0.75 | 0.73 | 0.70 | 0.65 | 0.45 |

[9] *Review of Economic Statistics*, 1919, p. 124.

It is obvious that a maximum correlation of 0.75 is obtained for a lag of either five or six months.

Whenever there is a high degree of correlation between two series for a definite amount of lag of a second series on a first, the first series may be looked upon as causing the movements in the second. Whether a causal relation exists or not, if the movements of the first persist in their tendency to precede the movements of the second, by a definite amount, the movements of the first are used in predicting the movements of the second.

It is advantageous in determining the lag or lead from a comparison of the plotted curves to have the data expressed in *standard units.* To express the data in standard units, we compute, first, the standard deviation for each series, and then divide the individual items of the series by the standard deviation for that series. If the original items are represented by $x_1$ and $x_2$, we may represent the corresponding items expressed in standard units by $X_1 = \frac{x_1}{\sigma_{x_1}}$ and $X_2 = \frac{x_2}{\sigma_{x_2}}\cdot$ In this notation, the correlation coefficient is simply $\frac{\Sigma X_1 X_2}{N}\cdot$

**137. Harvard forecasting sequence.**—The Harvard University Committee on Economic Research has made use of the lag in different series to determine a forecasting sequence. Their methods are described in a pamphlet "The Interpretation of the Index of General Business Conditions" by W. L. Crum, reprinted from *Review of Economic Statistics,* 1925, pp. 217–235, with revisions to October, 1926. Extracts from this pamphlet are reproduced here with the permission of the Harvard University Committee on Economic Research.

> . . . in the normal conditions which prevailed for many years prior to the World War, there was a fairly regular sequence in the cyclical fluctuations of those statistical series which represent respectively:
>
> A. Speculation.
> B. Business.
> C. Credit (Money).
>
> Any one cycle, as an isolated episode in business history, may have peculiar characteristics which distinguish it from other cycles. It may

have a peculiar time duration, a peculiar amplitude (intensity of the cyclical deviation from normal), and a peculiar form (reflected by the steepness of the recovery from depression and ascent into prosperity, and by the swiftness of the subsequent decline at the time of crisis). Moreover, the dissimilarities between individual cycles are so great that an average of several cycles may convey little reliable information about any particular cycle. In other words, as particular cycles differ from each other, an average may tell us little concerning the time-duration, amplitude, and form of a particular cycle.

On the other hand, the sequence between the cycles in speculation, business, and credit is a much more regular phenomenon and is much less subject to variation. This sequence affords, therefore, a valuable means of forecasting the movements of one type from a knowledge of the movements of the two other types. Such forecasting requires a careful interpretation of the economic significance of the situation reflected by the three types of movements and hence the sequence itself is not a sufficient means of forecasting.

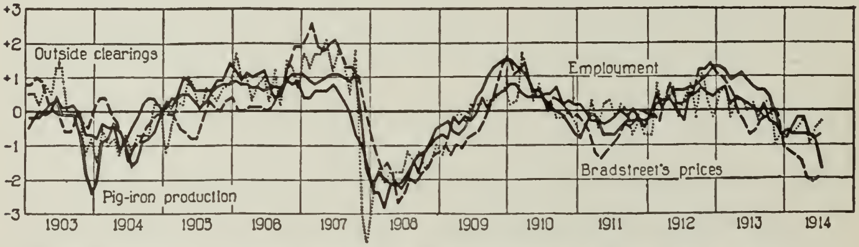


**Fig. 47.—Comparison of Cyclical Fluctuation in Bradstreet's Wholesale Commodity Price Index; Bank Clearings outside New York City, Pig-iron Production in the United States, and Employment: Monthly, 1903—July, 1914.**

Figure 47 shows that the cycles for the curves which represent the activity of business and the price level are synchronous: a particular phase of a given cycle for one curve occurs at the same time as that phase for another curve. This condition of concurrent fluctuation is found for a large number of series which represent *business* conditions.

Fig. 48 compares the curves of industrial stock prices and New York City clearings, and here also the fluctuations are concurrent. If, however, one of these curves is compared with one of the curves of Fig. 47, a marked difference in the timing of fluctuations appears. Thus, in Fig. 49, the movements in industrial stock prices tend to precede corresponding movements in commodity prices. In like manner, cyclical fluctuations in New York City clearings precede corresponding fluctuations for any of the curves of Fig. 47. As industrial stock prices and New York City clearings reflect the price level and volume of activity, respectively, in *speculation,* the general statement appears that *cyclical fluctuations in speculation precede corresponding fluctuations in business.*

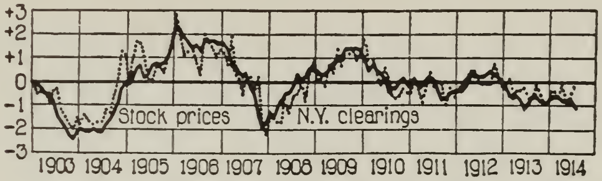


Fig. 48.—Comparison of Cyclical Fluctuations in the Dow-Jones Price Index of Industrial Stocks, and Bank Clearings in New York City: Monthly, 1903—July, 1914.

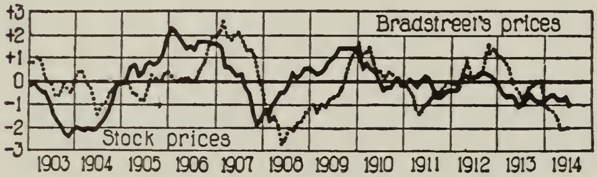


Fig. 49.—Comparison of Cyclical Fluctuations in the Dow-Jones Price Index of Industrial Stocks, and Bradstreet's Wholesale Commodity Price Index: Monthly, 1903—July, 1914.

Figures 50 and 51 make analogous comparisons for curves reflecting the "price level" and the volume of activity in credit . . . The conclusion is that *cyclical fluctuations in credit succeed corresponding fluctuations in business.*

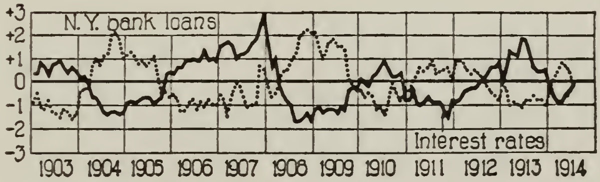


Fig. 50.—Comparison of Cyclical Fluctuations in the Rate on 60–90 Day Prime Commercial Paper in New York City, and the Loans of New York Clearing House Banks: Monthly, 1903—July, 1914.

Hence there is established a *sequence of cyclical* movements: a particular phase of a given cycle appears, first in speculation, then in business, and finally in credit. After the examination and analysis of a large number of series of economic data it is possible, by a process of selection, to form one group each for speculation, business, and credit. A composite index of the curves in each group can then be found, and the three resulting curves constitute an index chart which exhibits the three trains of cycles. Figure 52 is constructed on this plan, and clearly exhibits the sequence in the movements of speculation, business, and money (credits).

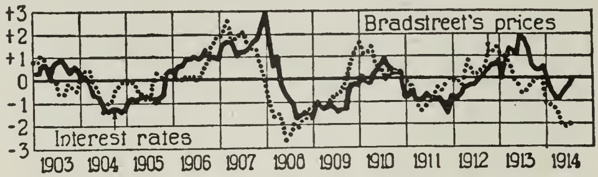


**Fig. 51.—Comparison of Cyclical Fluctuations in the Rate on 60–90 Day Prime Commercial Paper in New York City, and Bradstreet's Wholesale Commodity Price Index: Monthly, 1903—July, 1914.**

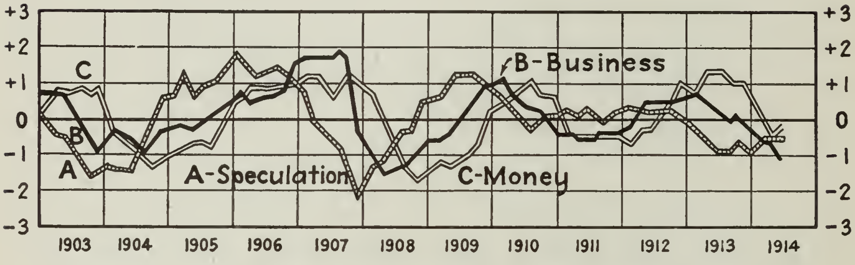
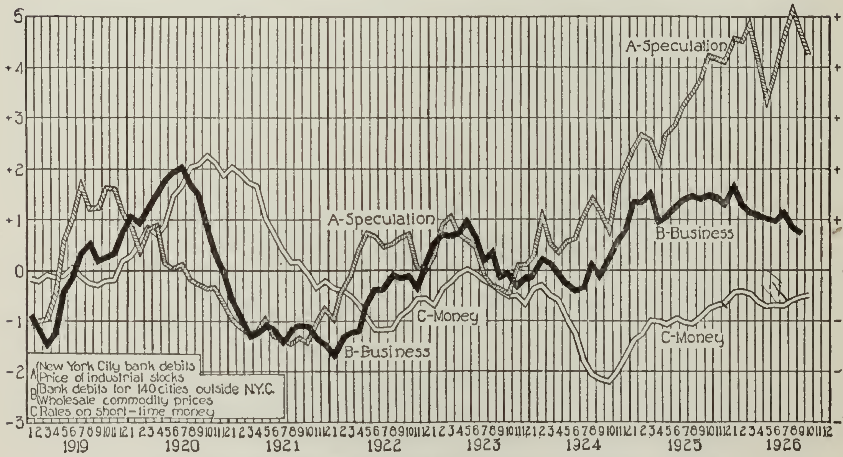


**Fig. 52.—The Harvard Index Chart; Bi-monthly, 1903—June, 1914.**



• The curves from 1924 on are based upon revisions described in this REVIEW, April, 1926, pp. 64–68.

**Fig. 53.—The Index of General Business Conditions: Monthly, 1919–26.**

After the war, . . . , the movements which were familiar before the war were resumed; and, although striking differences in trend and some alterations in the relative positions of the curves appeared, the main aspects of the cyclical fluctuations which we knew before the war again came to the fore. Figure 53 shows the index of general business conditions from 1919 to 1927.

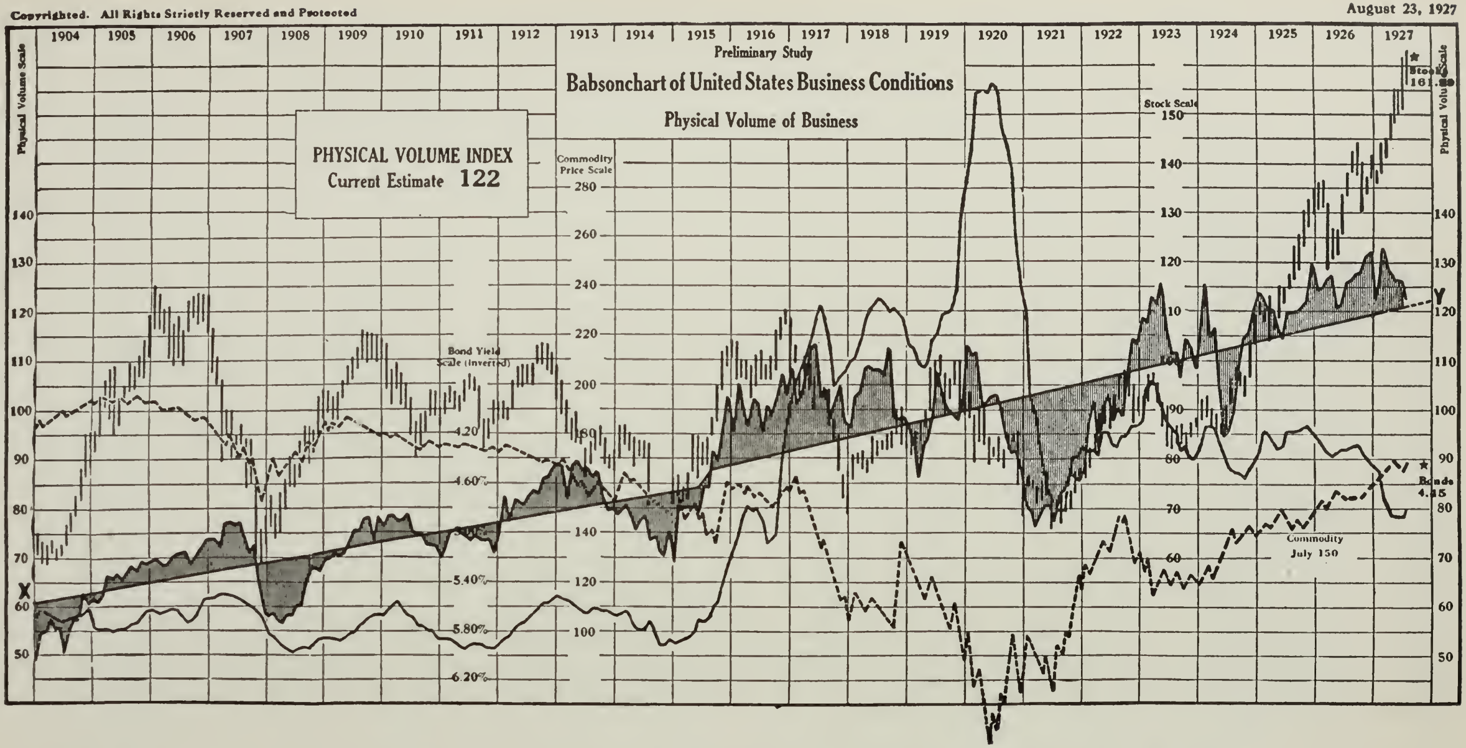
### E. Babsonchart.

**138. Babsonchart.**—We give on p. 228 by permission a copy and explanation of the Babsonchart of United States Business Conditions.

This study of the Babsonchart is based on 45 different subjects, carefully weighted and carried back over a twenty-three year period. The subjects used are now divided into four main groups: (1) Basic Materials: consisting of the production of Pig Iron, Copper, Bituminous Coal, Anthracite Coal, Petroleum and Lumber. (2) Agricultural Marketings: namely, Cattle, Hog and Sheep Receipts, Wool Receipts, Cotton into Sight, Corn, Wheat, and Oats Receipts. (3) Manufactures: represented by Cotton and Wool Consumption, Production of Beef, Pork and Lard, Lamb and Mutton, Steel Ingots, Flour, Sugar Meltings, Lumber Shipments and New Building. (4) Transportation and Trade: including ton-miles of Class I roads, Imports and Exports.

"The *X*-*Y* line represents the country's net gain or growth. Based on the economic theory that 'action and reaction are equal' when the two factors of time and intensity are multiplied to form an area, the sums of the areas above and below said line *X*-*Y* must, over sufficiently long periods of time be equal, provided enough subjects are included and properly weighted and combined. In this chart the *X*-*Y* line is calculated by the method of moments; first for the period 1904 through 1915 inclusive, and then for the period 1915 through 1926 inclusive, the two lines being joined by a similarly computed trend line for the years 1915, 1916, which are the years of transition to the new level of activity brought about by the stimulation of the World War.

"The high points of the stock market usually come in the early part of the over-expansion areas. The low points have come in the early part of the depression areas."



The vertical lines show the monthly range of the average of 40 stocks (20 rails and 20 industrials). Use scale figures under 1923.

- - - The dotted line is a record of the average yield of twenty active bonds. The figures under 1911 (with scale inverted) refer to this line.

* Star marks last quotation before going to press.

——The unbroken line is a record of the average wholesale prices of commodities, excluding foodstuffs. Scale figures under 1923.

Fig. 54.

The formula used in the calculation of this index is Fisher's Ideal. In its form to give price, it is

$$\sqrt{\frac{\Sigma P_1Q_o}{\Sigma P_oQ_o}\,\frac{\Sigma P_1Q_1}{\Sigma P_oQ_1}};$$

to give volume it is

$$\sqrt{\frac{\Sigma P_oQ_1}{\Sigma P_oQ_o}\,\frac{\Sigma P_1Q_1}{\Sigma P_1Q_o}};$$

where $P_1$ = current price; $Q_1$ = current quality; $P_o$ = base price (average monthly price during the years 1919–1923); $Q_o$ = base quality similarly divided.

### Exercises.

1. Plot the data in table 80.

2. Compute the standard deviation for each of the series in table 80.

3. Express the data in table 80 in standard units.

4. Compute the coefficient of correlation for the data in table 80, ssuming the lag of interest rates to be one year.

5. Plot the price per bushel received by producers, Dec. 1, for corn as given by the Yearbook of the Department of Agriculture. Cover the period from 1870 to the present. (See table XV, appendix.)

## Chapter XII

## BINOMIAL DISTRIBUTION

**139. Probability.**[1]—If an event can happen in $h$ ways and fail in $f$ ways, and each of these ways is equally likely to occur, then $\frac{h}{h+f}$ is said to be the *probability* that the event will happen and $\frac{f}{h+f}$ is said to be the probability that the event will fail to take place.

In applying this definition, one must be sure that the events are equally likely to occur. In a throw of a die, any one of the six sides is equally likely to fall uppermost. In a toss of a coin, heads or tails are equally likely to appear. When an individual goes to bat in a ball game, either one of two events may happen: either he will make a safe hit or he will not. These events are not equally likely. Here the individuality of the batter enters and, if he is unknown, we have nothing on which to make an estimate.

In throwing a die, there are six equally likely events. One face is just as likely to turn up as any other. The ace can happen in one of these ways. Hence the probability of throwing an ace with a single throw is $\frac{1}{6}$. When one selects a card at random from a pack of playing cards, there are 52 cards each equally likely to be drawn, and of these 52 cards 13 are hearts. Then the probability of drawing a heart in a single draw is $\frac{13}{52}$.

If an event is certain to happen, there are no ways in which it can fail, and $f = 0$.
Hence

$$\frac{h}{h+f} = 1$$

Thus, if an event is certain to happen, its probability is 1.

[1] Adapted from Lovitt and Holtzclaw, "Mathematics of Business," D. Appleton and Co., New York, pp. 124–125.

If an event is certain to fail, its probability is 0, for then

$$h = 0 \text{ and } \frac{h}{h+f} = 0$$

Thus 1 is the symbol for certainty and 0 the symbol for impossibility. In every other case the probability that an event will happen is between 0 and 1, since $h < h + f$.

If $p$ is the probability that an event will happen, the probability $q$ that the event will not happen is $1 - p$, for

$$\frac{h}{h+f} + \frac{f}{h+f} = 1$$

whence

$$\frac{f}{h+f} = 1 - \frac{h}{h+f}$$

or

$$q = 1 - p$$

There is a class of events for which it is impossible to enumerate all of the equally likely ways in which one of them may happen or fail to happen. For example, suppose we desire to know the probability that a man aged 40 will live one more year. We have no way of enumerating the equally likely ways in which this event may happen or fail. What we do is this. From a large group of representative men aged 40 who have been observed in the past, it has been found that, out of 68,297 men alive on their 40th birthday, 67,583 survived to celebrate their 41st birthday. Then the fraction $\frac{67{,}583}{68{,}297} = 0.98954$ is taken to be the measure of the probability that any one man aged 40 will reach the age of 41. This is sometimes known as *statistical probability*.

**140. A fundamental principle.**—If a first act can be done in $r$ ways, and after it has been done, a second act can be performed in $s$ ways, then the number of ways in which the two acts can be performed, in the order named, is $rs$.

For, since the second act can be performed in $s$ ways, then for each way of doing the first act there are $s$ ways of performing both acts. Then, for the $r$ ways of doing the first act, there are $rs$ ways of performing both acts.

Thus, if there are 3 roads from $A$ to $B$, and 4 roads from $B$ to $C$, there are $3 \times 4 = 12$ different roads leading from $A$ to $C$.

**141. Independent events.**—Two events are said to be *independent* when the happening or failing to happen of one of them has nothing to do with the happening or failing to happen of the other.

Thus the turning up of heads on one penny has nothing to do with the turning up of heads on a second penny or the turning up of heads on the same penny on a second throw.

Theorem I: *The probability that both of two independent events will happen is the product of their separate probabilities.*

Let us suppose that the separate probabilities of the two events are $p_1$ and $p_2$, where

$$p_1 = \frac{h_1}{h_1 + f_1} \text{ and } p_2 = \frac{h_2}{h_2 + f_2}$$

Then, by the above fundamental principle, the total number of ways in which the two events can happen or fail to happen is $(h_1 + f_1)(h_2 + f_2)$, while the number of ways in which both events can happen is $h_1h_2$. Then, by the definition of probability, the probability $p$ that both events will happen is

$$\frac{h_1h_2}{(h_1 + f_1)(h_2 + f_2)} = p_1p_2$$

Therefore

$$p = p_1p_2$$

Thus, if the probability that heads will turn up on one coin is $\frac{1}{2}$ and the probability that heads will turn up on a second coin is $\frac{1}{2}$, then the probability that heads will turn up on both coins is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. In general, if $p_1, p_2, \ldots, p_n$ are the simple probabilities of a number of separate independent events, then the probability $p$ that all of the events will happen is the product $p_1, p_2, \ldots, p_n$ of their separate possibilities. That is

$$p = p_1p_2 \cdots p_n$$

Corollary I: *If $p_1, p_2, \ldots, p_n$ are the simple probabilities of a number of separate independent events, then the probability that the first r of the events will happen and the rest fail is*

$$p_1p_2 \cdots p_r(1 - p_{r+1}) \cdots (1 - p_n)$$

Thus, when three coins are tossed, the probability that heads will turn up on the first coin and not turn up on the other two is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

The probability, when two dice are rolled, that the ace will turn up on the first but not on the second is $\frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$.

**142. Mutually exclusive events.**—If two or more events are so related that the occurrence of any one of them on a given occasion precludes the possibility of the occurrence of any of the others on that occasion, then the events are said to be *mutually exclusive.*

Thus on a single throw with a die, if the ace turns up, then none of the other faces can turn up on that throw. If on a single toss of two coins the combination, heads on the first and tails on the second, turns up, one cannot on that throw have the combination tails on the first and heads on the second.

Theorem II: *The probability that some one of a set of mutually exclusive events will happen is the sum of their respective probabilities.*

For, let $h_1, h_2, \ldots, h_r$ represent the number of ways of happening of the first, second, . . . , $r$th of the mutually exclusive events. Then $h_1 + h_2 + \cdots + h_r$ represents the total number of ways in which all of these events can happen. Let $n$ represent the total number of ways in which all of these events can happen and fail, all of these ways being equally likely. Then the respective chances of happening of the single events are

$$\frac{h_1}{n}, \frac{h_2}{n}, \ldots, \frac{h_r}{n}$$

The chance that some one of the $r$ events will happen is

$$\frac{h_1 + h_2 + \cdots + h_r}{n} = \frac{h_1}{n} + \frac{h_2}{n} + \cdots + \frac{h_r}{n}$$

That is, the chance that some one of the $r$ events will happen is equal to the sum of their individual chances of happening.

Thus, the probability of obtaining either a 4 or a 5 in a single throw with a die is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

**143. Permutations.**—If from $n$ things $r$ are chosen, and arranged in order, each arrangement is called a *permutation.* Two permutations are different if they contain different things, or the same things, arranged in a different order. Thus *abc* and *abd* are different permutations, since the second contains the letter $d$, which is not in the first. Also *abc* and *bac* as permutations are different, since there is a different arrangement of the same letters.

The number of permutations that can be made from $n$ different things, $a_1, a_2, \ldots, a_n$, taken $r$ at a time, is represented by the symbol ${}_nP_r$ and is the same as the number of ways in which $n$ different things can be placed in $r$ boxes, one thing in each box. There are $n$ choices of things to put in the first box, leaving $n - 1$ things from which to choose something to put in the second box. Thus the first two boxes can be filled in $n(n - 1)$ ways. This leaves $n - 2$ things from which to choose something to put in the third box. Thus, the first three boxes can be filled in $n(n - 1)(n - 2)$ ways. Continuing in this way, we see that the number of ways in which $r$ boxes can be filled is

$${}_nP_r = n(n - 1)(n - 2) \cdots (n - r + 1)$$

If

$$r = n,$$

we have

$${}_nP_n = n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1 = n!$$

**144. Combinations.**—A *combination* is a set of things without reference to the order in which they are arranged. Thus, *abc* and *bac* are the same combination. In order that two combinations may be different, one must contain an element not in the other. Thus, *abc* and *abd* are different combinations. The number of combinations of $n$ things, taken $r$ at a time, is represented by the symbol ${}_nC_r$.

The number of combinations of $n$ things, taken $r$ at a time, can be determined from the following considerations.

Take any one combination of $r$ things. These $r$ things can be arranged into $r!$ permutations. Hence the total number of permutations which can be formed from ${}_nC_r$ combinations is

$$r!\,{}_nC_r = {}_nP_r$$

whence

$${}_nC_r = \frac{{}_nP_r}{r!}$$

That is

$${}_nC_r = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}$$

**145. Repeated trials.**—Given the probability $p$ that an event will happen in a single trial, we desire the probability that the event will happen exactly $r$ times in $n$ trials. Let $q = 1 - p$ represent the probability that the event will not happen in a single trial.

By the corollary to Theorem I, the probability that the event will happen $r$ times and fail $n - r$ times, in a specified order, is $p^rq^{n-r}$. But the number of ways in which the order can be specified is the number of combinations one can make from $n$ things taken $r$ at a time, or ${}_nC_r$. These combinations are equally likely and mutually exclusive. Hence, the probability that the event will happen exactly $r$ times in $n$ trials is

$${}_nC_rp^rq^{n-r}$$

The probability that an event will happen *at least* $r$ times in $n$ trials is the sum of the probabilities that it will happen $n, n-1, n-2, \cdots, r$ times, or

$$p^n + {}_nC_{n-1}p^{n-1}q + {}_nC_{n-2}p^{n-2}q^2 + \cdots + {}_nC_rp^rq^{n-r}$$

Thus, the probability that one will obtain an ace exactly 5 times in 7 throws of a single die, or exactly 5 times in a single throw with 7 dice, is

$${}_7C_5(\tfrac{1}{6})^5(\tfrac{5}{6})^2 = \frac{7\cdot6\cdot5\cdot4\cdot3}{1\cdot2\cdot3\cdot4\cdot5}\,\frac{5^2}{6^7} = \frac{175}{110{,}592}$$

The probability that one will obtain an ace *at least* 5 times in 7 throws is

$$(\tfrac{1}{6})^7 + 7(\tfrac{1}{6})^6(\tfrac{5}{6}) + \frac{7\cdot6}{1\cdot2}(\tfrac{1}{6})^5(\tfrac{5}{6})^2 = \frac{561}{331{,}776}$$

### Exercises.

1. Find the probability, in drawing a card from a pack of 52 playing cards, that it will be (a) an ace; (b) a diamond; (c) a face card; (d) not a face card. *Ans.* $\frac{1}{13}$, $\frac{1}{4}$, $\frac{4}{13}$, $\frac{9}{13}$.

2. If two pennies are tossed simultaneously, what is the probability that one will obtain (a) two heads; (b) two tails; (c) one head and one tail? *Ans.* $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{2}$.

3. Find the probability that on a single throw with two dice the sum will be (a) seven; (b) eight; (c) nine; (d) ten; (e) eleven. *Ans.* $\frac{1}{6}$, $\frac{5}{36}$, $\frac{1}{9}$, $\frac{1}{12}$, $\frac{1}{18}$

4. What is the probability of throwing 2 aces in a single throw with 2 dice? Three aces in a single throw with 3 dice? *Ans.* $\frac{1}{36}$, $\frac{1}{216}$

5. What is the probability of throwing either an ace or a deuce with a single throw of a die? *Ans.* $\frac{1}{3}$

6. What is the probability of throwing either the combination 6, 4 or the combination 5, 5 in a single throw with two dice? *Ans.* $\frac{1}{12}$

7. What is the probability of drawing either the ace or the king of hearts in a single draw from a pack of 52 cards? *Ans.* $\frac{1}{26}$

8. What is the probability of drawing either an ace or a king in a single draw from a pack of 52 cards? *Ans.* $\frac{2}{13}$

9. What is the probability that the ace will turn up exactly 3 times in 6 throws of a single die? *Ans.* $\frac{2500}{46656}$

10. What is the probability that the ace will turn up at least 3 times in 6 throws of a single die? *Ans.* $\frac{2906}{46656}$

11. In five throws with a single die what is the probability of throwing an ace (a) exactly twice; (b) at least twice? *Ans.* $\frac{1250}{7776}$, $\frac{1526}{7776}$

12. In five throws of a single coin what is the probability of throwing exactly 2 heads? At least 2 heads? *Ans.* $\frac{10}{32}$, $\frac{26}{32}$

13. When five coins are tossed, what is the probability of 2 heads and 3 tails turning up? *Ans.* $\frac{10}{32}$

14. If 10 coins are tossed, find the probability of turning up exactly

| *Ans.* | *Ans.* | *Ans.* |
|---|---|---|
| (a) 2 heads, $\frac{45}{1024}$; | (b) 3 heads, $\frac{120}{1024}$; | (c) 4 heads, $\frac{210}{1024}$; |
| (d) 5 heads, $\frac{252}{1024}$; | (e) 6 heads, $\frac{210}{1024}$; | (f) 7 heads, $\frac{120}{1024}$; |
| (g) 8 heads, $\frac{45}{1024}$; | (h) 9 heads, $\frac{10}{1024}$; | (i) 10 heads, $\frac{1}{1024}$; |

(j) either 2 or 3 heads; (k) either 4, 5, or 6 heads;

(l) not less than 8 heads; (m) not less than 7 heads.

**146. The binomial distribution.**—If we toss two coins, we may obtain two heads, two tails, or one head and one tail. Let us use $H$ to represent heads and $T$ to represent tails and the numbers (1) and (2) to distinguish the coins. Then, if we toss two coins, we may record the expected results as follows:

| (1) (2) | or | (1) (2) | or | (1) (2) |
|---|---|---|---|---|
| *H H* | | *H T* | | *T T* |
| | | *T H* | | |

More briefly, this may be written

$$H^2 + 2HT + T^2 \equiv (H + T)^2$$

If we toss 3 coins, we may expect

| (1) (2) (3) | or | (1) (2) (3) | or | (1) (2) (3) | or | (1) (2) (3) |
|---|---|---|---|---|---|---|
| *H H H* | | *H H T* | | *H T T* | | *T T T* |
| | | *H T H* | | *T H T* | | |
| | | *T H H* | | *T T H* | | |

This may be written

$$H^3 + 3H^2T + 3HT^2 + T^3 \equiv (H + T)^3$$

If we toss 4 coins, we may expect

| (1) (2) (3) (4) | or | (1) (2) (3) (4) | or | (1) (2) (3) (4) | or | (1) (2) (3) (4) | or | (1) (2) (3) (4) |
|---|---|---|---|---|---|---|---|---|
| *H H H H* | | *H H H T* | | *H H T T* | | *H T T T* | | *T T T T* |
| | | *H H T H* | | *T H H T* | | *T H T T* | | |
| | | *H T H H* | | *T T H H* | | *T T H T* | | |
| | | *T H H H* | | *H T T H* | | *T T T H* | | |
| | | | | *H T H T* | | | | |
| | | | | *T H T H* | | | | |

This may be written

$$H^4 + 4H^3T + 6H^2T^2 + 4HT^3 + T^4 \equiv (H + T)^4$$

We observe that we have $6H^2T^2$ because there are 6 combinations of four things taken 2 at a time. That is, with 4 coins, there are 6 possible ways of obtaining the combination of 2 tails.

In general, if $n$ coins are thrown, the relative frequencies of 0, 1, 2, . . . tails are given by the terms of the binomial expansion $(H + T)^n$.

A simple rule can be given for the expansion of the binomial $(H + T)^n$.

1. The first term is $1 \cdot H^n$.
2. The exponent on $H$ decreases by one for each succeeding term.
3. The second term contains $T^1$. The exponent on $T$ increases by one for each succeeding term.
4. The last term is $1 \cdot T^n$.
5. The coefficient of the second term is $n$.
6. The coefficients of succeeding terms are obtained as follows:

   Multiply the coefficient of any term by the exponent of $H$ in that term and divide the result by one more than the exponent of $T$ in that term. The result is the coefficient for the next term.

In tossing two coins, the probability of obtaining 2 heads is $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$. The probability of obtaining one head and one tail is $2(\frac{1}{2})(\frac{1}{2}) = \frac{1}{2}$. The probability of obtaining two tails is $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$. These probabilities are given immediately by the terms of the expansion

$$(\tfrac{1}{2} + \tfrac{1}{2})^2 = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4}$$

If we toss 3 coins, the probabilities of obtaining 0, 1, 2, 3 tails are $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$. These probabilities are given immediately by the terms of the expansion

$$(\tfrac{1}{2} + \tfrac{1}{2})^3 = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8}$$

In general, the *probabilities* of 0, 1, 2, 3, . . . , $n$ tails in a single throw of $n$ coins are given by the successive terms in the expansion of $(\frac{1}{2} + \frac{1}{2})^n$, while the *frequencies* of 0, 1, 2, 3, . . . , $n$ tails in $N$ throws of $n$ coins are given by the successive terms in the expansion of $N(\frac{1}{2} + \frac{1}{2})^n$. Thus, the frequencies of 0, 1, 2 tails in 100 throws of 2 coins are given by

$$100(\tfrac{1}{2} + \tfrac{1}{2})^2 = 100(\tfrac{1}{2})^2 + 2 \cdot 100(\tfrac{1}{2})(\tfrac{1}{2}) + 100(\tfrac{1}{2})^2 = 25 + 50 + 25.$$

In general, if $p$ is the probability of success of an event (say, throw of an ace with a single die) and $q$ is the probability of the failure of the event, then the frequencies of 0, 1, 2, 3,

. . . successes in $N$ trials of $n$ independent events are given by the successive terms of the binomial expansion

$$N(q + p)^n$$

**Exercises.**

1. Compute to the nearest unit the terms of the binomial series $1{,}000(q + p)^{10}$ for values of $q$ from 0.1 to 0.9 by steps of 0.1.

2. Compute for $q = 0.9$, $p = 0.1$ the terms of $1{,}000(q + p)^n$ for $n = 2$, 3, 5.

3. Compute $1{,}000(\frac{1}{6} + \frac{5}{6})^n$ for $n = 2, 3, 4, 5$.

4. Actually throw 2 dice 1,000 times and count and tabulate the number of aces on each throw and compare with the computed frequencies.

5. Compute and plot $1{,}000(q + p)^n$ for $q = p = \frac{1}{2}$; $n = -2$.

**147. Comparison of actual and theoretical frequencies.**—Let us throw 6 pennies a total of 128 times. At each throw let us count the number of tails appearing and tabulate. This was done with the following result:

| number of tails: | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| frequency: | 1 | 12 | 30 | 37 | 33 | 12 | 3 |

The expected frequency is given by the terms of the binomial expansion $128(\frac{1}{2} + \frac{1}{2})^6$. We have

$$128(\tfrac{1}{2} + \tfrac{1}{2})^6 = 2 + 12 + 30 + 40 + 30 + 12 + 2$$
$$= 128\left(\frac{1 + 6 + 15 + 20 + 15 + 6 + 1}{\phantom{64}}\right)$$

If one plots the data for the actual frequencies, using number of tails as abscissa and frequencies as ordinates, one obtains a very good approximation to the bell-shaped symmetrical curve. The theoretical frequencies do give a symmetrical curve. If the reader will keep $p = q = \frac{1}{2}$ and plot the frequency curves for increasing values of $n$, he will find that the frequency polygons remain symmetrical and the number of sides of the polygon increase. As $n$ increases indefinitely, the polygon becomes a smooth curve, which is known as the normal probability curve, or normal frequency curve.

In many cases the frequency distribution of natural phenomena follows this same probability curve. Some people state that the reason why the frequencies for natural phenomena follow the normal frequency distribution is that each element in the determination of a characteristic has, like the coin, a 50-50 chance.

**148. Arithmetic average and standard deviation of a binomial distribution.**—Let us use as origin $X_o = 0$ successes. Consider the following table of successes:

| $X$<br>*No. of successes* | $f$<br>*Frequency* |
|---|---|
| 0 | $q^n$ |
| 1 | $npq^{n-1}$ |
| 2 | $\frac{n(n-1)}{1 \cdot 2} p^2 q^{n-2}$ |
| 3 | $\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} p^3 q^{n-3}$ |
| $\vdots$ | $\vdots$ |

| $fX$ | $fX^2$ |
|---|---|
| 0 | 0 |
| $npq^{n-1}$ | $npq^{n-1}$ |
| $n(n-1)p^2q^{n-2}$ | $2n(n-1)p^2q^{n-2}$ |
| $\frac{n(n-1)(n-2)}{1 \cdot 2} p^3 q^{n-3}$ | $\frac{3n(n-1)(n-2)}{1 \cdot 2} p^3 q^{n-3}$ |
| $\vdots$ | $\vdots$ |

We have

$$\Sigma f = N = (q + p)^n = 1$$

$$\begin{aligned}
\overline{X} &= \frac{\Sigma fX}{N} = \Sigma fX \\
&= np\left[q^{n-1} + (n-1)q^{n-2}p + \frac{(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2 + \cdots\right] \\
&= np(q + p)^{n-1} = np
\end{aligned}$$

$$\begin{aligned}
\sigma^2 &= \frac{\Sigma fX^2}{N} - \overline{X}^2 = \Sigma fX^2 - \overline{X}^2 \\
&= np\left[q^{n-1} + 2(n-1)q^{n-2}p + 3\frac{(n-1)(n-2)}{1 \cdot 2} q^{n-3}p^2 + \cdots\right] \\
&\qquad - n^2p^2 \\
&= np[(q + p)^{n-1} + (n-1)p(q + p)^{n-2}] - n^2p^2 \\
&= np[1 + (n-1)p] - n^2p^2 \\
&= np - np^2 = np(1 - p) = npq
\end{aligned}$$

$$\therefore \sigma = \sqrt{npq}$$

Let us compute the arithmetic average and standard deviation for the data given in article 147 for the throw of 6 pennies a total of 128 times.

For the theoretical frequencies we find

$$\overline{X} = np = 6 \times \tfrac{1}{2} = 3$$
$$\sigma = \sqrt{npq} = \sqrt{6 \times \tfrac{1}{2} \times \tfrac{1}{2}} = 1.2247$$

For the actual frequencies we find

| $X$ | $f$ | $fX$ | $fX^2$ |
|---|---|---|---|
| 0 | 1 | 0 | 0 |
| 1 | 12 | 12 | 12 |
| 2 | 30 | 60 | 120 |
| 3 | 37 | 111 | 333 |
| 4 | 33 | 132 | 528 |
| 5 | 12 | 60 | 300 |
| 6 | 3 | 18 | 108 |
| Totals | $N = 128$ | 393 | 1,401 |

$$\overline{X} = \frac{\Sigma fX}{N} = \tfrac{393}{128} = 3.007$$
$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \overline{X}^2} = \sqrt{\frac{1,401}{128} - (3.007)^2} = 1.38$$

Observe that, whether we use the theoretical or actual frequencies, the total range for $X$, which is 6 units, is more than 4 and less than 5 times the standard deviation.

**149. Probability of a deviation greater than a given multiple of the standard deviation.**—It can be demonstrated that, for the normal probability curve, one-half of the area under the curve is included between the ordinates drawn at $x = \overline{X} - 0.6745\sigma$ and $x = \overline{X} + 0.6745\sigma$. With respect to the variates under observation in a given sample, this means that the chances are even that a variate, chosen at random from the sample, will have for its measure a value within the stated range $\overline{X} \pm 0.6745\sigma$.

The following table gives the chances that there shall occur a deviation from the estimated average greater than the stated multiple of the standard deviation.

| *Range* | *Approximate Chances* |
|---|---|
| $\pm 0.6745\sigma$ | 1 to 1 |
| $\pm\sigma$ | 2 to 1 |
| $\pm 2\sigma$ | 21 to 1 |
| $\pm 3\sigma$ | 369 to 1 |
| $\pm 4\sigma$ | 15,772 to 1 |

The meaning can be illustrated as follows. Suppose one obtains the arithmetic average height of 200 Scandinavian workmen in a given factory, and computes the standard deviation for the group. Then, if a workman's height is measured and found to differ from this arithmetic average by more than three times the standard deviation. there is a very high probability that this workman does not belong to the group of 200 who were measured, or, indeed, to the group of Scandinavian workmen of which the 200 were a sample.

Indeed, it has been found in practice that, for a symmetrical or a moderately skewed distribution, the total range seldom exceeds six times the standard deviation. We have already called attention to the fact that, for both the actual and theoretical frequencies of the throw of six pennies a total of 128 times, the total range for $x$ is less than 5 times the standard deviation.

*Illustration.*—Determine whether it is to be expected that one will obtain 2,164 tails in 4,096 throws of a single coin.

$$\overline{X} = np = 4{,}096 \times \tfrac{1}{2} = 2{,}048$$

$$\sigma = \sqrt{npq} = \sqrt{4{,}096 \times \tfrac{1}{2} \times \tfrac{1}{2}} = 32$$

$$3\sigma = 96$$

$$2{,}048 + 96 = 2{,}144 < 2{,}164$$

The chances are very much against obtaining as many as 2,164 tails and hence 2,164 tails are not to be expected.

**Exercises.**

1. Determine whether it is to be expected that one will obtain

| | *Ans.* |
|---|---|
| (a) 562 tails in 1,024 throws of a coin. | No. |
| (b) 1,070 heads in 2,304 throws of a coin. | No. |
| (c) 3,300 tails in 6,400 throws of a coin. | Yes. |

(d) 5,450 heads in 10,816 throws of a coin. *Ans.* Yes.

(e) 1,600 appearances of an ace with 8,820 throws of a single die. *Ans.* Yes.

(f) 8,700 appearances of an ace with 52,020 throws of a single die. *Ans.* Yes.

(g) in 350 times out of 5,508 throws with two dice the sum 11 on the upper face. *Ans.* Yes.

(h) 230 appearances of the sum 7 or 11 in 1,134 throws with two dice. *Ans.* Yes.

(i) 130 trumps in 432 deals of a single card. *Ans.* Yes.

(j) 34 deaths due to typhoid in a population of 100,000 when the death rate in the general population is 25 per 100,000. *Ans.* Yes.

2. In 300 times at bat a man makes 36 home runs. Is it to be expected that in the next 25 times at bat the batter

(a) will make no home runs? *Ans.* Yes.

(b) will make 7 home runs? *Ans.* Yes.

(c) will make 10 home runs? *Ans.* No.

3. In shooting at clay pigeons a man's past record is 81 out of 100 broken. Is it to be expected that out of 76

(a) he will break as many as 73? *Ans.* No.

(b) he will break no more than 48? *Ans.* No.

4. A boy is winding peach buds in a nursery. In the past 36 per cent of his buds grew.

(a) What is the smallest number out of 1,000 which is likely to grow? *Ans.* 314.

(b) What is the greatest number likely to grow? *Ans.* 406.

(c) If less than 314 grew, what would be your conclusion?

(d) If more than 406 grew, what would be your conclusion?

5. Two dice are thrown 216 times. Compute the theoretical frequencies of the sums on the upper faces.

| *Sum* | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 6 | 12 | 18 | 24 | 30 | 36 | 30 | 24 | 18 | 12 | 6 |

6. Would you be willing to bet 10 to 1 that an opponent could not throw the sum 7 with two dice at least 23 times in a hundred throws with two dice? *Ans.* Yes.

7. Would you be willing to bet 15 to 1 that an opponent could not throw the double ace at least 4 times in 36 throws with two dice? *Ans.* Yes.

8. Three dice are thrown 216 times. Compute the theoretical frequencies of the sums on the upper faces.

9. The standing of a baseball team for the season is 0.36. Is it to be expected that in their next 25 games *Ans.*

(a) they should win 19 or more? No.
(b) they should win none? No.
(c) they should win one? No.
(d) they should win two? Yes.

10. In the following instances, is the variation from the average such as to justify one in constructing a theory, based on this variation, as to the causes of the variation? In other words, is there a significant difference in the averages?

(a) The annual average for the period 1906–1910 of the number of deaths per 100,000 of population due to typhoid fever as given by the 1919 report of the census bureau in its bulletin on Mortality Statistics is as follows:

| *Registration States* | $\overline{X} = 24.6$ |
|---|---|
| Colorado | 43.3 |
| Massachusetts | 13.7 |

*Range* $= \overline{X} \pm 3\sigma = 9.72$ to $39.48$; $\sigma = 4.96$.

(b) Death rate per 100,000 due to scarlet fever:

| | *Annual Average*<br>1906–1910 |
|---|---|
| *Registration States* | $\overline{X} = 10.5$ |
| California | 3.4 |
| Colorado | 21.0 |

*Range* $= \overline{X} \pm 3\sigma = 0.78$ to $20.22$; $\sigma = 3.24$.

(c) Tuberculosis (all forms), death rate per 100,000:

| *Registration States* | $\overline{X} = 163.5$ |
|---|---|
| California | 210.4 |
| Colorado | 244.2 |
| Michigan | 99.7 |
| N. Y., Bronx district | 445.8 |
| Scranton, Penn | 97.4 |

*Range* $= 125.16$ to $201.84$; $\sigma = 12.78$.

(d) Tuberculosis (all forms):

| *Registration Cities in Non-Registration States* | $\overline{X} = 196.5$ |
|---|---|
| Seattle | 110.3 |
| Portland | 110.8 |
| Omaha | 127.1 |
| Kansas City, Mo | 163.4 |

*Range* $= 196.5 \pm 42 = 154.5$ to $238.5$; $\sigma = 14$.

(e) Malaria:

| (I) *Registration States* | $\overline{X} = 1.6$ |
|---|---|
| California | 3.5 |
| Colorado | 0.6 |

*Range* = −2.195 to 5.395; $\sigma$ = 1.265.

| II *Registration Cities in Non-Registration States* | $\overline{X} = 7.7$ |
|---|---|
| New Orleans | 12.9 |
| Chicago | 0.5 |
| Milwaukee | 0.1 |
| Memphis | 104.7 |

*Range* = −0.6 to 16.0; $\sigma$ = 2.77.

(f) Suicides (1922):

| (I) *Registration States* | $\overline{X} = 11.9$ |
|---|---|
| California | 25.3 |
| Colorado | 18.0 |
| North Carolina | 4.2 |

*Range* = 1.7 to 22.1; $\sigma$ = 3.4.

| (II) *Registration Cities* | $\overline{X} = 14.3$ |
|---|---|
| Cambridge, Mass | 6.3 |
| Wilmington, Del | 7.8 |
| New York City | 13.8 |
| Chicago | 14.4 |
| Denver | 23.5 |
| Omaha | 26.4 |
| San Francisco | 30.6 |
| Los Angeles | 30.7 |

*Range* = 2.9 to 25.7; $\sigma$ = 3.8.

(g) Automobile Accidents (1922):

| (I) *Registration States* | $\overline{X} = 12.5$ |
|---|---|
| California | 26.0 |
| Colorado | 16.3 |
| Mississippi | 3.4 |
| Massachusetts | 12.5 |

*Range* = 2 to 23; $\sigma$ = 3.5.

| (II) *Registration Cities* | $\overline{X} = 16.8$ |
|---|---|
| Chicago | 22 |
| New York City | 896 |
| Philadelphia | 267 |
| Lowell, Mass | 5.2 |

*Range* = 4.5 to 29.1; $\sigma$ = 4.1.

(h) Deaths from old age (1910):

| *Registration States* | $\overline{X} = 25.3$ |
|---|---|
| Colorado | 26.9 |
| Maine | 77.3 |
| Vermont | 51.7 |
| Ohio | 19.0 |

*Range* = 10.3 to 40.3; $\sigma$ = 5.0.

11. The deflection for a ray of light, grazing the surface of the sun, should be

| | |
|---|---|
| 1.″75 | *Einstein's theory* |
| 0.″87 | *Newton's theory.* |

From photographs taken in May, 1919, at Sobral in North Brazil and at the Isle of Principe in the Gulf of Guinea, West Africa, the deflections with their standard deviations were found to be

| | |
|---|---|
| *Sobral*. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 1.″98, $\sigma$ = 0.″18 |
| *Principe*. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 1.″61, $\sigma$ = 0.″45 |

The value of the material obtained at Principe (due to cloudy weather) cannot be put higher than about one-sixth that at Sobral. (*Source: Eddington, "Space, Time, and Gravitation," Cambridge University Press, 1920.*)

Do the results found at the Isle of Principe rule out the Newtonian theory? *Ans.* No.

Do the results found at Sobral rule out the Newtonian theory? *Ans.* Yes.

Is Einstein's theory consistent with the results found at both places? *Ans.* Yes.

**150. Application in insurance.**—According to the American Experience Mortality Table, out of 100,000 living at age ten, 749 die within the year. On this basis, the probability of death within a year of a child aged 10 is 0.00749. If an insurance company had 10,000 lives aged 10 insured, the number of expected deaths is 74.9. The standard deviation from this number is

$$\sigma = \sqrt{10{,}000 \times 0.00749 \times 0.99251} = 8.62$$

The probable error in the number of expected deaths is

$$0.6745 \times 8.62 = 5.81$$

The maximum error in the number of expected deaths is

$$3\sigma = 25.86$$

If the company computed the premium on the basis of a death rate of $74.9 + 3\sigma = 100.76$, then the company would be practically certain of experiencing no loss on the 10,000 insured lives.

If the company computed the premium on the basis of $74.9 + 5.81 = 80.7$ deaths, then the chances are even that the company will experience no loss on the 10,000 insured lives.

### Exercises.

1. When the probability of death within a year is 0.007 on 200 insured lives, what should be the assumed death rate for the insurance company to have an even chance of experiencing no loss on the business? *Ans.* 0.0110

2. By the United States Life Tables for 1910, the rate of mortality per thousand for both sexes in the original registration states at various ages is as tabulated below. From a population of 10,000, compute the most probable number of deaths and the maximum deviation to be expected under normal circumstances.

| | | *Ans.* |
|---|---|---|
| *Age* | *Rate* | *Maximum Deviation* |
| 20 | 4.68 | 20.4 |
| 30 | 6.51 | 24 |
| 40 | 9.39 | 29 |
| 50 | 14.37 | 36 |
| 60 | 28.58 | 51 |
| 70 | 59.52 | 70 |

# Chapter XIII

## SOME CHARACTERISTIC CURVES

**151. Straight line.**—A general form for the equation of a straight line is

$$y = mx + b$$

The graph for the straight line whose equation is $y = \frac{x}{2} - 2$ is shown in figure 55.

Two points determine a straight line. From the equation, when $x = 0$, $y = -2$. This determines the point $A$. When $y = 0$, $x = 4$. This determines the point $B$.

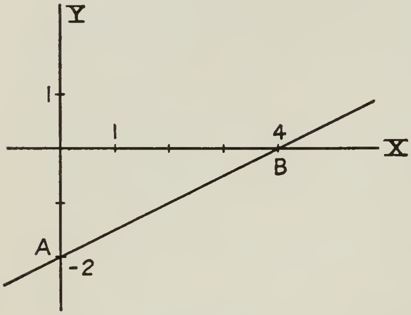


**Fig. 55.**

For the straight line $y = \frac{x}{2} - 2$ let us tabulate a few values of $y$ for values of $x$ equally spaced.

**Table 81.—Difference Table for $y = x/2 - 2$.**

| $x$ | $y$ | $\Delta y$ = *First Differences of y Values* |
|---|---|---|
| 0 | −2 | |
| | | 0.5 |
| 1 | −1.5 | |
| | | 0.5 |
| 2 | −1 | |
| | | 0.5 |
| 3 | −0.5 | |
| | | 0.5 |
| 4 | 0 | |
| | | 0.5 |
| 5 | 0.5 | |
| | | 0.5 |
| 6 | 1 | |

Compute the differences of the adjacent $y$-values. These differences in the present instance are all equal to 0.5. These values are termed the first differences of $y$. Sometimes the symbol $\Delta y$ is used to represent these first differences.

It is a characteristic property of a straight line that the first differences of the values of $y$, for values of $x$ equally spaced, are constant.

**152. Parabola.**—A general form for the equation of a parabola is

$$y = ax^2 + bx + c$$

For all values of $a$, $b$, and $c$, the curve has the same general appearance, which is illustrated by the graph of $y = x^2$ in figure 56.

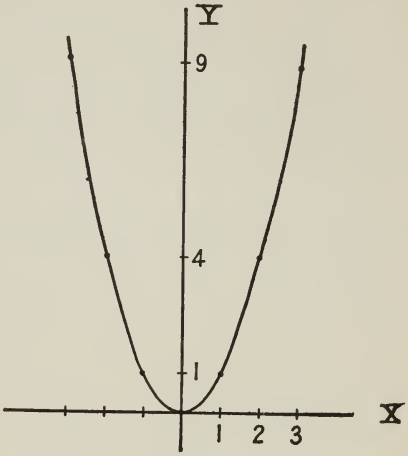


Fig. 56.

Table 82.—Difference Table for $y = x^2$.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ |
|---|---|---|---|
| −3 | 9 | | |
| | | −5 | |
| −2 | 4 | | 2 |
| | | −3 | |
| −1 | 1 | | 2 |
| | | −1 | |
| 0 | 0 | | 2 |
| | | 1 | |
| 1 | 1 | | 2 |
| | | 3 | |
| 2 | 4 | | 2 |
| | | 5 | |
| 3 | 9 | | |

Table 82 shows that the first differences of $y(\Delta y)$ are not constant. The differences of the numbers in the column headed $\Delta y$ are called the second differences of $y$. The symbol $\Delta^2 y$ is used to represent these second differences. The first differences of the second differences are called third differences, and are represented by $\Delta^3 y$. A corresponding notation is used for higher differences.

From table 82 it is seen that for this parabola the second differences of $y$ are constant, namely 2.

It is a characteristic property of a parabola $y = ax^2 + bx + c$ that the second differences of $y$, for values of $x$ differing by unity, are constant.

**153. Rectangular hyperbola.**—The most useful form of the equation of this curve for statistical purposes is

$$xy = c \text{ (a constant).}$$

The general form is displayed by the graph in fig. 57 of the equation

$$xy = 1$$

| $x$ | $y$ |
|---|---|
| $\frac{1}{4}$ | 4 |
| $\frac{1}{3}$ | 3 |
| $\frac{2}{3}$ | $\frac{3}{2}$ |
| 1 | 1 |
| 2 | $\frac{1}{2}$ |
| 3 | $\frac{1}{3}$ |
| 4 | $\frac{1}{4}$ |
| $-\frac{1}{3}$ | $-3$ |
| $-\frac{2}{3}$ | $-\frac{3}{2}$ |
| $-1$ | $-1$ |
| $-2$ | $-\frac{1}{2}$ |
| $-3$ | $-\frac{1}{3}$ |
| $-4$ | $-\frac{1}{4}$ |

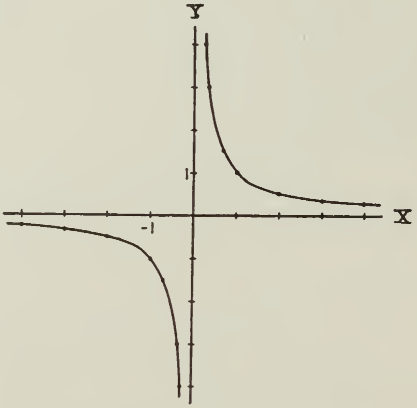


Fig. 57.

If in place of $y$ we put $\frac{1}{Y}$, our equation becomes

$$Y = \frac{x}{c}$$

which is linear in $Y$ and $x$. Hence, if for values of $x$ differing by unity we compute the first differences of the reciprocals of the corresponding ordinates, these differences will be constant.

**154. Third degree curve.**—The most general equation of a curve of third degree is

$$y = ax^3 + bx^2 + cx + d$$

For all values of *a*, *b*, *c*, and *d*, these curves have the same general appearance, which is illustrated by the curve (fig. 58)

$$y = x^3 - 4x$$

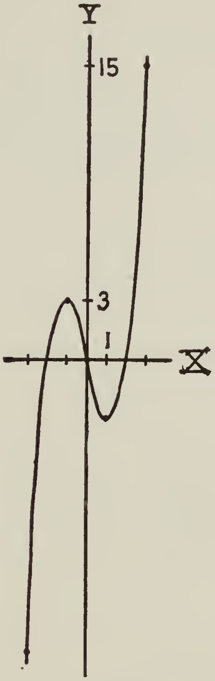


Fig. 58.

**Table 83.—Difference Table for $y = x^3 - 4x$.**

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| −3 | −15 | | | |
| | | 15 | | |
| −2 | 0 | | −12 | |
| | | 3 | | 6 |
| −1 | 3 | | − 6 | |
| | | −3 | | 6 |
| 0 | 0 | | 0 | |
| | | −3 | | 6 |
| 1 | − 3 | | 6 | |
| | | 3 | | 6 |
| 2 | 0 | | 12 | |
| | | 15 | | |
| 3 | 15 | | | |

For every curve of this type the third differences ($\Delta^3 y$) are constant

**155.** ***N*-th degree curve.**—The most general equation of a curve of degree $n$ is

$$y = a_o x^n + a_1 x^{n-1} + \cdot \cdot \cdot + a_{n-1}x + a_n$$

All curves of this type have the same general appearance, cutting the $x$-axis in $n$ points, real or imaginary. Figure 59 gives illustrations of curves of degree four, five, and six.

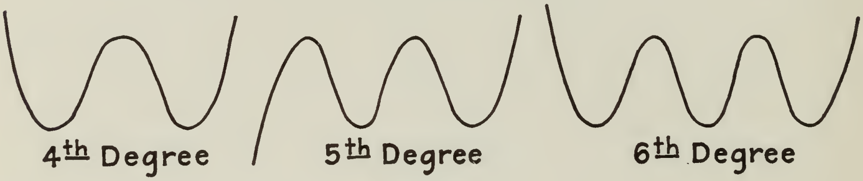


Fig. 59.

For every curve of this type, the $n$-th differences of the ordinates, for equal spaced values of $x$, are constant.

**156. Exponential curves.**—The set of curves

$$y = a \cdot 10^{bx}$$

for all values of $a$ and $b$ have the same general appearance, which is illustrated by (fig. 60)

$$y = (1.5)^x$$

| $x$ | $y$ |
|---|---|
| 0 | 1 |
| 1 | 1.5 |
| 2 | 2.25 |
| 3 | 3.37 |
| 4 | 5.06 |
| 5 | 7.59 |
| −1 | 0.67 |
| −2 | 0.44 |
| −3 | 0.30 |

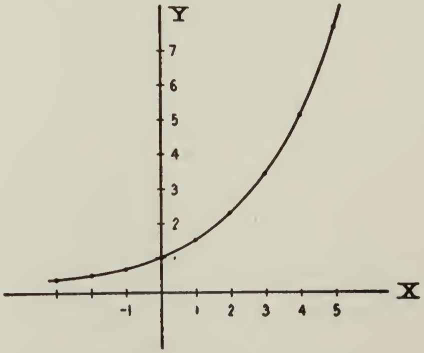


Fig. 60.

This curve is sometimes known as the compound interest curve (see fig. 60).

Take the logarithms of both sides of the general equation. We find

$$\log y = \log a + bx$$

This equation is linear in log $y$ and $x$. Hence the first differences of log $y$, for equal spaced values of $x$, are constant. This is illustrated by table 84.

Table 84.—Difference Table for $\log y = \log 1.5 + x$.

| $x$........ | $-3$ | $-2$ | $-1$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|
| $y$........ | .29629 | .44444 | .66666 | 1 | 1.5 | 2.25 | 3.375 | 5.0625 | 7.5937 |
| log $y$..... | $-0.52827$ | $-0.35218$ | $-0.17609$ | 0 | 0.17609 | 0.35218 | 0.52827 | 0.70436 | 0.88045 |
| Δ · log $y$....... | 0.17609 | 0.17609 | 0.17609 | 0.17609 | 0.17609 | 0.17609 | 0.17609 | 0.17609 | |

**157. Probability curve.**—The normal probability curve

$$y = \frac{k}{\sqrt{\pi}} e^{-k^2x^2} \quad (e = 2.71828+)$$

is one curve from the set

$$y = A \cdot 10^{-Bx^2}$$

For all positive values of $A$ and $B$, these curves have the same general shape, which is illustrated by

$$y = 2^{-x^2}$$

| $x$ | $y$ |
|---|---|
| $-2$ | $\frac{1}{16}$ |
| $-1$ | $\frac{1}{2}$ |
| 0 | 1 |
| 1 | $\frac{1}{2}$ |
| 2 | $\frac{1}{16}$ |

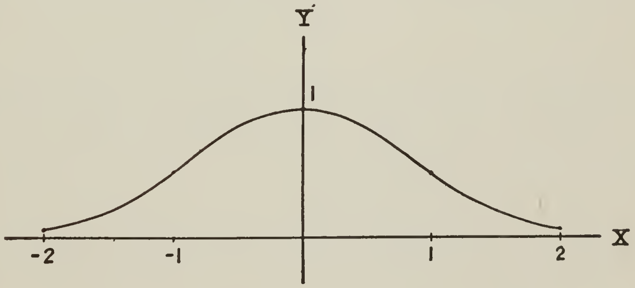


Fig. 61.

This is the symmetrical bell-shaped curve. Take the logarithms of both sides of the general equation. We find

$$\log y = \log A - Bx^2$$

If in place of log $y$ we put $Y$, we have

$$Y = \log A - Bx^2$$

which is a parabola. Hence, we see that for values of $x$ equally spaced, the second differences of log $y$ are constant. This is illustrated by the table 85 for $y = 2^{-x^2}$.

Table 85.—Difference Table for log $y$ where $y = 2^{-x^2}$.

| $x$.......... | $-2$ | $-1$ | $0$ | $1$ | $2$ |
|---|---|---|---|---|---|
| $y$.......... | 0.0625 | 0.5 | 1 | 0.5 | 0.0625 |
| log $y$....... | $-1.20412$ | $-0.30103$ | 0 | $-0.30103$ | $-1.20412$ |
| $\Delta \cdot \log y$......... | $-0.90309$ | $-0.30103$ | $+0.30103$ | $+0.90309$ | |
| $\Delta^2 \cdot \log y$.... | | $-0.60206$ | $-0.60206$ | $-0.60206$ | |

The equation[1]

$$y = \frac{A}{B + Cx^2}$$

for positive values of $A$, $B$, and $C$ has the same general appearance as the probability curve. For this curve the second differences of the reciprocals of the ordinates are constant.

The equation

$$y = \frac{A}{e^x + e^{-x}}$$

for positive values of $A$ has also the same general form as the probability curve and on occasion is used in place of either of the above.

**158. Miscellaneous curves.**—An equation of the form

$$y = \frac{x}{a + bx + cx^2}$$

for positive values of $x$ may represent a *skew symmetric* curve. To illustrate, we show, for positive values of $x$, the plot (fig. 62) of $y = \dfrac{4x}{x^2 - 4x + 9}$

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $y$ | 0 | $\frac{2}{3}$ | $\frac{8}{5}$ | 2 | $\frac{16}{9}$ | $\frac{10}{7}$ | $\frac{8}{7}$ | $\frac{14}{15}$ |

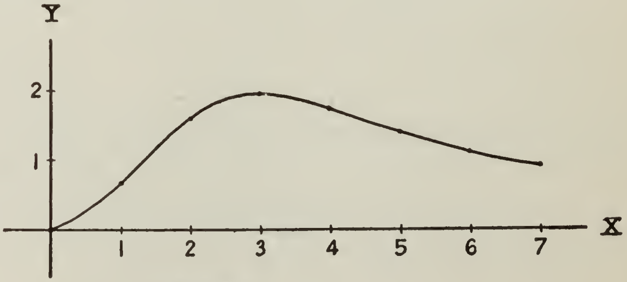


Fig. 62.

[1] For an example of the use of this curve see Dowling and Turneaure, "Analytic Geometry," p. 191, Henry Holt and Company, New York.

For other examples consult almost any text on Strength of Materials.

The equation[2]

$$y = 10^{0.81-0.36x+0.03x^2}$$

represents a U-shaped curve. The graph is shown in fig. 63.

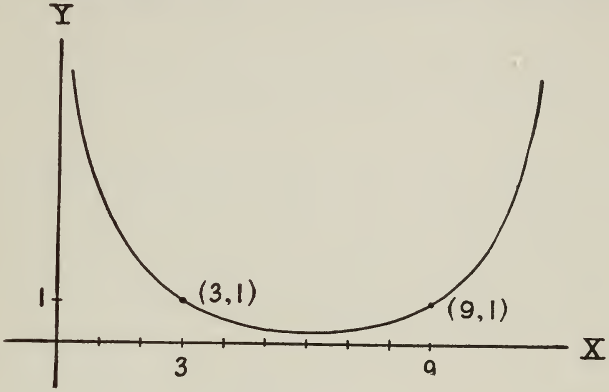


Fig. 63.

**Exercises.**

Plot the following curves.

1. $y = 2x - 1; x - 3; 3x - 2; \frac{1}{2}x + 3; -2x + 1.$
2. $y = x^2, 2x^2, \frac{1}{2}x^2, x^2 - 1, x^2 - 4, x^2 - 9.$
3. $xy = 1, 2, 4, 5, -1, -2, -6.$
4. $y = x^2 - 7x + 12, x^2 - 5x + 6$
5. $y = (x - 1)(x - 2)(x - 3).$
   $y = (x - 1)(x^2 + x + 1).$
   $y = (x + 1)(x^2 - 4x + 3).$
6. $y = 0.5^x, 2^x, (1.2)^x, (1.3)^x, (1.6)^x.$
7. $y = 10^{0.54-0.24x+0.02x2}$
   $y = 10^{-0.54+0.24x-0.02x2}$
8. $y = \dfrac{x}{0.2 - 0.1x + 0.05x^2}, y = \dfrac{x}{0.2 - 0.1x + 0.1x^2}$
   $y = \dfrac{4x}{x^2 - 4x + 9}, y = \dfrac{3x}{x^2 + x + 1}$
9. $y = 2{\cdot}10^{-0.01x^2}, 10^{-0.1x^2}, 10^{-0.5x^2}.$
10. $y = \dfrac{1}{1 + x^2}, \dfrac{1}{1 + 2x^2}, \dfrac{1}{1 + 5x^2}, \dfrac{4}{1 + x^2}$

**159. Determination of curve type.**—Given a series of statistical data, the question arises as to what kind of curve will best fit the data.

[2] Many other interesting examples of empirical curves and their uses can be found in Running, T. R., "Empirical Formulas." N. Y., Wiley, 1917.

*First method—Observation.*—One who is acquainted with various types of curves will plot the data. The plotted points may appear to the unaided eye to arrange themselves in a general way along one of the typical curves. The general equation of this type is assumed, and one proceeds to determine the parameters in such a way as to obtain as good a fit as possible.

*Second method.*—If values of $x$ are equally spaced, compute the first, second, third, etc., differences of $y$. If first differences are constant or nearly so, assume that the data fit a straight line. If second differences are more nearly constant than any of the other differences, assume that the data fit a parabola. If the $n$-th differences are constant or nearly so, assume that the data fit a curve of degree $n$. If none of these differences seems to be constant, difference the reciprocal of the $y$ values. If these differences are nearly constant, assume a hyperbolic form. If second differences of the reciprocals are nearly constant, assume $y = \frac{1}{(a + bx + cx^2)}$. If first differences of log $y$ are constant, assume $y = A \cdot 10^{Bx}$. If second differences of log $y$ are constant, assume $y = A \cdot e^{-Bx^2}$.

## CHAPTER XIV

# CURVE FITTING

**160. Introduction.**—In practice, the relation between quantities is usually not known in advance, but is to be found, if possible, from pairs of numerical values of the quantities obtained while conducting an experiment, making a set of observations, or collecting statistical data on any matter whatsoever.

Having obtained a set of values $Y_1, Y_2, \ldots, Y_n$ of one variable corresponding to the values $x_1, x_2, \ldots, x_n$ of a related variable, two problems present themselves. *First,* what is the most suitable equation $y = f(x, a, b, c, \cdots)$ to choose to represent the empirical data? In making this choice we desire that the equation shall be simple and that it shall give a close fit to the data. *Second,* how shall we determine the parameters $a, b, c, \ldots$ in the selected equation so that the mathematical curve shall give the best fit to the empirical data? By *best fit* we usually mean that for the given set of values of $x$, namely $x_1, x_2, \ldots, x_n$, the arithmetic average of the sum of the differences between the computed values $y_1, y_2, \ldots, y_n$ and the observed values $Y_1, Y_2, \ldots, Y_n$ of $y$, shall be zero.

**161. Selection of equation.**—This matter has been discussed in the last article of the preceding chapter. Graphically, or by computation of the proper differences, we determine whether the data can be properly fitted by one of the following types of curves:

1. $y = mx + b$ (*straight line*)
2. $y = a + bx + cx^2$
   $x = a + by + cy^2$ (*parabola*)
3. $y = ax^n$ (*parabolic in form*)
4. $xy = c$ (*hyperbola*)

5. $y = ax^{-n}$ (*hyperbolic in form*)

6. $y = a \cdot 10^{bx}$ (*exponential*)

7. $y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$ (*normal probability*)

8. $y = a \cdot 10^{-bx^2}$ (*symmetric frequency*)

   $y = \frac{a}{b + cx^2}$

9. $y = \frac{x}{a + bx + cx^2}$ $(b^2 - 4ac < 0)$ (*skew symmetric*)

10. $y = K \cdot 10^{a - bx + cx^2}$ (*u-shaped*)

If the plotted data do not fit any of the types given above, a general method of procedure is to assume an equation of the type

$$y = a_o + a_1x + a_2x^2 + \cdots a_nx^n$$

By taking $n$ one less than the given number of pairs of values we can determine $a_o$, $a_1$, . . . , $a_n$ so as to make this curve pass through all of the plotted points given by the empirical data

### Determination of the Parameters.

**162. Graphical.**—If the plotted points appear to be approximately on a straight line, by the aid of a transparent ruler, draw a straight line through the points in such a way that there are about as many points on one side of it as on the other. Then the distance from the origin to the point where the straight line cuts the $y$-axis is the measure of the parameter $b$ in the equation $y = mx + b$. Select two points $(x_1, y_1)(x_2, y_2)$ on the straight line, preferably several units apart. Measure $y_2 - y_1$ and $x_2 - x_1$. Then $\frac{(y_2 - y_1)}{(x_2 - x_1)}$ is the value of the parameter $m$, which is called the slope of the straight line.

If the plotted points appear to be on a curve of the type $xy = c$, plot $x$ and $Y = \frac{1}{y}$. These values of $Y$ and $x$ will be approximately on a straight line $Y = \frac{x}{c}$. This straight line should go through the origin. The coördinates $(x_o, Y_o)$ of any point will give the slope $\frac{1}{c} = \frac{Y_o}{x_o}$. Hence $c = \frac{x_o}{Y_o} = x_o y_o$. That is, if the product $xy$ of the observed values is nearly constant, a proper value of $c$ can be obtained graphically by measuring the slope of the line $Y = \frac{x}{c}$.

In case the plotted points appear to be on one of the parabolic or hyperbolic curves of the group

$$y = ax^m$$

take the logarithm (base 10) of both sides:

$$\log y = m \log x + \log a$$

and then substitute $X$ for $\log x$, $Y$ for $\log y$, $b$ for $\log a$, so that the equation becomes

$$Y = mX + b$$

Thus, we see that if we plot $\log x$ and $\log y$ instead of the original data, the plotted points will be on a straight line whose slope $m$, and $y$-intercept $b$, can be found graphically. Then from the equation $b = \log a$, we can find $a$ and hence the parameters $a$ and $m$ in the equation $y = ax^m$ are known.

In like manner one can obtain graphically the parameters in some of the other forms.

**163. Selected points.**—If the plotted points appear to be on a straight line, select from the plotted points two points $(x_1, y_1)$ $(x_2, y_2)$ such that the straight line through them appears to be a fair fit. Assume that the equation of the straight line through these points is of the form

(1) $$y = mx + b$$

Then we have the two equations

$$y_1 = mx_1 + b$$
$$y_2 = mx_2 + b$$

from which to determine $m$ and $b$.

The choice of points here is arbitrary and the resulting straight line very often is not a good fit for the original data.

The extension to other curve types is obvious.

*Example:*

| $x$ | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| $y$ | 48 | 66 | 80 | 98 |

The straight line through (20, 48) (50, 98) is obtained by substituting these values of $x$ and $y$ in (1). We find

$$48 = 20m + b$$
$$98 = 50m + b$$

Solving these equations for $m$ and $b$, we find

$$m = \frac{5}{3},\ b = 14\frac{2}{3}$$

Therefore

$$y = \frac{5}{3}x + 14\frac{2}{3}$$

In decimal notation we write this as follows:

$$y = 1.7x + 14.7$$

The straight line through

| | |
|---|---|
| (20, 48)(40, 80) is | $y = 1.6x + 16$ |
| (30, 66)(40, 80) is | $y = 1.4x + 24$ |
| (30, 66)(50, 98) is | $y = 1.6x + 18$ |

We do not try the other combinations, for, from a graph, the corresponding lines obviously are not a good fit.

**164. Moments.**—If the plotted points *appear* to be upon a straight line, a parabola, or a curve of the $n$-th degree, the corresponding equation is assumed and we proceed to determine the coefficients by a method which is illustrated in the following examples:

*Example I:* Let the observed values of $x$ and $y$ be

| $x$ | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| $y$ | 48 | 66 | 80 | 98 |

The plotted points appear to be on a straight line.

Furthermore, the first differences are more nearly constant than any other differences. Hence we assume

$$y = mx + b$$

We determine the parameters $m$ and $b$ in such a way that the zero-th and first moments for the ordinates of the straight line shall be equal to the corresponding moments for the observed ordinates.

The zero-th moment of the ordinates of the straight line is simply the sum of the ordinates for $x = 20, 30, 40, 50$ or

$$(20m + b) + (30m + b) + (40m + b) + (50m + b) = 140m + 4b$$

The zero-th moment of the observed ordinates is their sum, or

$$48 + 66 + 80 + 98 = 292$$

whence

$$140m + 4b = 292 \tag{2}$$

The first moment of the ordinates of the straight line is obtained by multiplying each ordinate by the corresponding value of $x$. We have

$$20(20m + b) + 30(30m + b) + 40(40m + b) + 50(50m + b) \\ \equiv 5{,}400m + 140b$$

The first moment of the observed ordinates is

$$20(48) + 30(66) + 40(80) + 50(98) = 11{,}040$$

Whence

$$5{,}400m + 140b = 11{,}040 \tag{3}$$

In practice, the simplest way to obtain these two equations is as follows: in the equation $y = mx + b$, replace $x$ and $y$ by their observed values. We have

$$\begin{aligned} 20m + b &= 48 \\ 30m + b &= 66 \\ 40m + b &= 80 \\ 50m + b &= 98 \end{aligned}$$

Multiply each equation by the coefficient of $b$ in that equation. Add the four equations. This gives equation (2). Multiply each equation by the coefficient of $m$ in that equation. Add the four equations. This gives equation (3). Solving equations (2) and (3), we find

$$m = 1.64,\ b = 15.6$$

whence

$$y = 1.64x + 15.6$$

Values of $y$ computed from this equation for the given values of $x$ are

| $x$ | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| $y$ | 48.4 | 64.8 | 82.2 | 97.6 |

*Example II:* Let the observed values of $x$ and $y$ be

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ | 3.2 | 2.1 | 2.9 | 5.8 | 11.3 |

The plotted points appear to be on a parabola. Furthermore, the second differences are more nearly constant than any of the other differences. Hence, we assume that

$$y = ax^2 + bx + c$$

We compute the parameters $a$, $b$, $c$ so that the zero-th, first and second moments of the observed values of $y$ shall be equal to the corresponding moments of the values of $y$ as computed from the assumed equation. For the zero-th moment we have

$$\begin{aligned} 3.2 + 2.1 + 2.9 + 5.8 + 11.3 &= (a + b + c) + (4a + 2b + c) \\ &+ (9a + 3b + c) + (16a + 4b + c) + (25a + 5b + c) \end{aligned}$$

or

$$55a + 15b + 5c = 25.3 \tag{4}$$

For the first moment we have

$$\begin{aligned} &(a + b + c) + 2(4a + 2b + c) + 3(9a + 3b + c) + 4(16a + 4b + c) \\ &+ 5(25a + 5b + c) = 3.2 + 2(2.1) + 3(2.9) + 4(5.8) + 5(11.3) \end{aligned}$$

or

$$225a + 55b + 15c = 95.8 \tag{5}$$

For the second moment we have

$$\begin{aligned} &(a + b + c) + 4(4a + 2b + c) + 9(9a + 3b + c) + 16(16a + 4b + c) \\ &+ 25(25a + 5b + c) = 3.2 + 4(2.1) + 9(2.9) + 16(5.8) + 25(11.3) \end{aligned}$$

or

$$979a + 225b + 55c = 413 \tag{6}$$

1t should now be obvious that, when the assumed equation is linear in the parameters, the moment equations can be obtained in the following systematic manner. In the assumed equation replace $x$ and $y$ by their observed values. In the present example we have

$$\begin{aligned} a + b + c &= 3.2 \\ 4a + 2b + c &= 2.1 \\ 9a + 3b + c &= 2.9 \\ 16a + 4b + c &= 5.8 \\ 25a + 5b + c &= 11.3 \end{aligned}$$

Multiply each equation by the coefficient of $c$ in that equation. Add the five equations. We obtain equation (4). Multiply each equation by the coefficient of $b$ in that equation.

Add the five equations. We obtain equation (5). Multiply each equation by the coefficient of $a$ in that equation. Add the five equations. We obtain equation (6). Solving equations (4), (5), (6), we find

$$a = 1.09, b = -4.55, c = 6.72$$

whence

$$y = 1.09x^2 - 4.55x + 6.72$$

**165. Method of averages.**—Let us suppose that there are to be determined $p$ parameters which enter linearly in the assumed equation, and that there are given $n$ pairs of values of the variables $x$ and $y$. Substitution of these $n$ pairs of values of $x$ and $y$ in the assumed equation will give $n$ equations for the determination of the $p$ parameters. Divide these $n$ equations into $p$ groups with, as nearly as possible, the same number of equations in each group. Add the equations in each group. We thus obtain $p$ equations for the determination of the $p$ parameters. The groupings are arbitrary and different groupings give different results. Let us illustrate this method by the data

| $x$ | 20 | 30 | 40 | 50 |
|---|---|---|---|---|
| $y$ | 48 | 66 | 80 | 98 |

Our equations are

$$\begin{aligned} 20m + b &= 48 \\ 30m + b &= 66 \\ 40m + b &= 80 \\ 50m + b &= 98 \end{aligned}$$

Adding the first two and then the last two, we obtain

$$\begin{aligned} 50m + 2b &= 114 \\ 90m + 2b &= 178 \end{aligned}$$

Solving for $m$ and $b$, we find

$$m = 1.6, b = 17$$

and hence

$$y = 1.6x + 17$$

By the method of moments we obtained $y = 1.64x + 15.6$. If we pair the first and third, and the second and fourth, we find

$$y = 1.8x + 12$$

**166. Equations not linear in parameters.**—In some cases these equations, by a proper transformation, can be reduced to equations which are linear in the parameters.

*Case I:* $y = ax^n$. Take the logarithms of both sides.

We have

$$\log y = \log a + n \log x$$

Make the substitutions

$$Y = \log y,\ A = \log a,\ X = \log x$$

Then

$$Y = A + nX \tag{7}$$

which is linear in the parameters $A$ and $n$. Having determined $A$, we find $a$ from the equation

$$A = \log a$$

*Example:* Let $x$ and $y$ be given as in the following table:

| $x$ | $y$ | $X = \log x$ | $Y = \log y$ |
|---|---|---|---|
| 2 | 6.0 | 0.3010 | 0.7782 |
| 4 | 24.6 | 0.6021 | 1.3909 |
| 8 | 70.8 | 0.9031 | 1.8500 |
| 16 | 338.8 | 1.2041 | 2.5299 |

Compute $X = \log x$ and $Y = \log y$ and place in parallel columns as shown in the table. The points $(X, Y)$ should lie approximately on a straight line. Substitute these values of $X$ and $Y$ in (7) and find $A$ and $n$ by the method of moments. We obtain

$$n = 1.914,\ A = 0.1970$$

Hence

$$a = 1.574$$

Therefore

$$y = 1.574x^{1.914}$$

*Case II:*

$$y = a.10^{bx}$$

Take the logarithm of both sides. We have

$$\log y = \log a + bx$$

Make the substitutions

$$\log y = Y,\ \log a = A$$

Then

$$Y = A + bx$$

which is linear in the parameters $A$ and $b$.

*Example:* The following table[1] gives the corresponding values of $x$ and $y$ where $x$ is the angle of contact between a belt and pulley expressed in radians and $y$ is the pull in lbs. required to raise a given weight.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ | 1.57 | 2.09 | 2.62 | 3.14 | 3.66 | 4.19 | 4.71 |
| $y$ | 5.62 | 6.93 | 8.52 | 10.50 | 12.90 | 15.96 | 19.67 |
| $Y = \log y$ | .7497 | .8407 | .9304 | 1.0212 | 1.1106 | 1.2030 | 1.2938 |

Solving by the method of averages, grouping the first four and the last three, we find

$$b = 0.173,\ A = 0.4782$$

Whence

$$a = 3.007$$

Therefore

$$y = 3.007 \times 10^{0.173x}$$

**Exercises.**[2]

1. The annexed tables give series of values of effort $E$, and load $R$, observed in testing a crane. Determine a relation of the form $E = mR + b$ connecting $E$ and $R$.

| | | | | | | | *Ans.* |
|---|---|---|---|---|---|---|---|
| (a) | $R$ | 10 | 20 | 30 | 40 | 50 | $m = 0.498$ |
| | $E$ | 6 | 10.8 | 16.1 | 20.8 | 26 | $b = 0.96$ |
| (b) | $R$ | 9 | 18 | 27 | 36 | 45 | $m = 0.337$ |
| | $E$ | 5 | 7.8 | 11.1 | 14.2 | 17 | $b = 1.90$ |
| (c) | $R$ | 10 | 20 | 30 | 40 | 50 | $m = 0.33$ |
| | $E$ | 4 | 7 | 11 | 14 | 17 | $b = 0.70$ |
| (d) | $R$ | 4 | 8 | 12 | 16 | 20 | $m = 0.2625$ |
| | $E$ | 2 | 3.1 | 4 | 5.2 | 6.2 | $b = 0.95$ |

2. A wire under tension is found by experiment to stretch an amount $l$, in thousandths of an inch, under a tension $T$, in pounds, as follows:

| | | | | | |
|---|---|---|---|---|---|
| $T$ | 3 | 6 | 9 | 12 | 15 |
| $l$ | 1.1 | 2.1 | 3.2 | 4.2 | 5 |

Find a relation of the form $l = mT$ (Hooke's law) which best represents these results. *Ans.* $l = 0.34T$.

[1] Running, T. R., "Empirical Formulas," p. 27 N. Y., Wiley, 1917.

[2] These exercises are taken from Kenyon-Lovitt, "Mathematics for Agriculture and General Science," Chapter XI. Macmillan Co., N. Y., 1918.

Reprinted by permission of the publishers.

3. The readings of a standard gas-meter $S$ and those of a meter $T$, being tested on the same pipe line, were found to be

| | | | | |
|---|---|---|---|---|
| $S$ | 3,000 | 3,510 | 4,022 | 4,533 |
| $T$ | 0 | 500 | 1,000 | 1,500 |

Find a formula of the type $T = mS + b$ which best represents these data. Interpret the values of $m$ and $b$.

4. The following table gives the density $\delta$ of liquid ammonia at various degrees centigrade. Determine a relation of the form $\delta = mt + b$

| | | | | |
|---|---|---|---|---|
| $t$ | 0 | 5 | 10 | 15 |
| $\delta$ | .6364 | .6298 | .6230 | .6160 |

*Ans.* $\delta = 0.6364 - 0.0014t$

5. The following table gives the specific heat $s$ of hot liquid ammonia at various degrees Fahrenheit. Find a relation of the form $s = mt + b$.

| | | | | | |
|---|---|---|---|---|---|
| $t$ | 5 | 10 | 15 | 20 | 25 |
| $s$ | 1.090 | 1.084 | 1.078 | 1.072 | 1.066 |

*Ans.* $s = 1.096 - 0.0012t$.

6. The distance $s$, in feet, passed over by a falling body in $t$ seconds is found by experiment to be

| | | | | | |
|---|---|---|---|---|---|
| $s$ | 0 | 5 | 16 | 35 | 65 |
| $t$ | 0 | 0.5 | 1 | 1.5 | 2 |

*Ans.* $s = 16.1t^2$

7. If a body slides down an inclined plane, the distance $s$, in feet, that it moves is connected with the time $t$, in seconds, after it starts by an equation of the form $s = kt^2$. Find the best value of $k$ consistent with the following data:

| | | | | | |
|---|---|---|---|---|---|
| $s$ | 2.6 | 10.1 | 23 | 40.8 | 63.7 |
| $t$ | 1 | 2 | 3 | 4 | 5 |

*Ans.* $k = 2.55$

8. If $\theta$ denotes the melting point (Centigrade) of an alloy of lead and zinc containing $x$ per cent of lead, it is found that

| | | | | | | |
|---|---|---|---|---|---|---|
| $x$ | 40 | 50 | 60 | 70 | 80 | 90 |
| $\theta$ | 186 | 205 | 226 | 250 | 276 | 304 |

Find a relation of the form $\theta = ax^2 + bx + c$.

*Ans.* $\theta = 0.0116x^2 + 0.8549x + 133.205$

9. The pressure $p$, measured in centimeters of mercury, and the volume $v$, measured in cubic centimeters, of a gas kept at constant temperature, were found to be as follows:

| | | | | | |
|---|---|---|---|---|---|
| $v$ | 145 | 155 | 165 | 178 | 191 |
| $p$ | 117.2 | 109.4 | 102.4 | 95 | 88.6 |

Determine a relation of the form $pv = k$.

*Ans.* $k = 16{,}936$.

10. A strong rubber band stretched under a pull of $p$ kg. shows an elongation of $E$ cm. Find a relation of the form $E = kp^n$.

| $p$ | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| $E$ | 0.1 | 0.3 | 0.6 | 0.9 | 1.3 | 1.7 | 2.2 | 2.7 | 3.3 | 3.9 |

*Ans.* $E = 0.3p^{1.6}$

11. The amount of water $A$, in cu. ft., that will flow per minute through 100 feet of pipe of diameter $d$, in inches, with an initial pressure of 50 lbs. per sq. in., is as follows:

| $d$ | 1 | 1.5 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|---|
| $A$ | 4.88 | 13.43 | 27.50 | 75.13 | 152.51 | 409.54 |

Find a relation of the form $A = kd^n$.

*Ans.* $A = 4.88d^{2.473}$

12. In testing a gas engine, corresponding values of the pressure $p$, measured in lbs. per sq. ft., and the volume $v$, in cubic feet, were obtained as follows:

| $v$ | 7.14 | 7.73 | 8.59 |
|---|---|---|---|
| $p$ | 54.6 | 50.7 | 45.9 |

Find a relation of the form $p = kv^n$. *Ans.* $p = 387.6v^{-0.938}$

13. Same as problem 12 but with the data

| $v$ | 6.27 | 5.34 | 3.15 |
|---|---|---|---|
| $p$ | 20.54 | 25.79 | 54.25 |

*Ans.* $pv^{1.41} = 273.5$

14. Given age in years and diameter in inches of a tree $1\frac{1}{2}$ feet from the ground as follows, plot the data and determine a relation of the form $y = kx^n$.

| *Age in Years* | 19 | 58 | 114 | 140 | 181 | 229 |
|---|---|---|---|---|---|---|
| *Diameter* | 3 | 7 | 13.2 | 17.9 | 24.5 | 33 |

*Ans.* $y = 6.96x^{1.03}$

15. Given age in years and height in feet of a tree as follows:

| *Age* | 13 | 34.4 | 50.5 | 218 | 247 |
|---|---|---|---|---|---|
| *Height* | 13.4 | 27.5 | 38.4 | 72.5 | 73 |

Plot the data and determine a relation of the form $y = kx^n$.

*Ans.* $y = 0.121x^{1.737}$

16. The intercollegiate track records for foot races are as follows, where $d$ means distance run, and $t$ the record time. Find a relation of the form $t = kd^n$.

| $d$, yds. | 100 | 220 | 440 | 880 | 1 mi | 2 mi |
|---|---|---|---|---|---|---|
| $t$ | $0{:}09\frac{4}{5}$ | $0{:}21\frac{1}{5}$ | 0:48 | $1{:}54\frac{4}{5}$ | $4{:}15\frac{2}{5}$ | $9{:}24\frac{2}{5}$ |

Data for the following problems are from the statistical abstract of the U. S., 1926, from the page indicated.

17. *p. 309.* The following data give the number of commercial failures in the years indicated for the United States. Fit (a) a straight line; (b) a parabola; and (c) a third degree curve.

**Commercial Failures in the U. S.**

| *Year* | *Number* | *Year* | *Number* |
|---|---|---|---|
| 1911 | 13,441 | 1919 | 6,451 |
| 1912 | 15,452 | 1920 | 8,881 |
| 1913 | 16,037 | 1921 | 19,652 |
| 1914 | 18,280 | 1922 | 23,676 |
| 1915 | 22,156 | 1923 | 18,718 |
| 1916 | 16,993 | 1924 | 20,615 |
| 1917 | 13,855 | 1925 | 21,214 |
| 1918 | 9,982 | 1926 | 21,773 |

18. *p. 733.* Fit a parabola to the following data on petroleum production. Quantities are in millions of barrels.

**Production of Crude Petroleum.**

| *Year* | *Quantity* | *Year* | *Quantity* |
|---|---|---|---|
| 1917 | 335 | 1922 | 558 |
| 1918 | 356 | 1923 | 732 |
| 1919 | 378 | 1924 | 714 |
| 1920 | 443 | 1925 | 764 |
| 1921 | 472 | 1926 | 767 |

Also fit a curve of the type $y = Ae^{Bx}$

19. The following data give the production and registration of motor vehicles. Beginning with 1920, data include production of plants located in Canada, making motor vehicles of U. S. design. Registration is for the U. S.

(a) Fit a parabola.
(b) Fit a curve of the type $y = ae^{bx}$.

**Production and Registration of Motor Vehicles in Millions. Statistical Abstract of the U. S., 1926, p. 369.**

| *Year* | *Production* | *Registration* | *Year* | *Production* | *Registration* |
|---|---|---|---|---|---|
| 1917 | 1.8 | 5.0 | 1922 | 2.6 | 12.2 |
| 1918 | 1.2 | 6.1 | 1923 | 4.1 | 15.1 |
| 1919 | 2.0 | 7.6 | 1924 | 3.6 | 17.6 |
| 1920 | 2.2 | 9.2 | 1925 | 4.3 | 19.9 |
| 1921 | 1.7 | 10.5 | 1926 | 4.4 | 22.0 |

20. *p. 374.* Automobile fatalities. Rate per estimated 100,000 population.

(a) Fit a straight line.
(b) Fit a parabola.

**Automobile Fatalities.**

| | *1920* | *1921* | *1922* | *1923* | *1924* | *1925* |
|---|---|---|---|---|---|---|
| *Registration States*............... | 10.3 | 11.4 | 12.5 | 14.8 | 15.6 | 16.9 |
| Cities...................... | 14.7 | 15.3 | 16.9 | 19.6 | 20.6 | 22.4 |
| Rural...................... | 6.2 | 7.6 | 8.4 | 10.4 | 11.1 | 12.1 |

21. *p. 267.* Building and loan associations: number, membership in millions, and assets in billions of dollars.

| *Year* | *Number* | *Members* | *Assets* | *Year* | *Number* | *Members* | *Assets* |
|---|---|---|---|---|---|---|---|
| 1916 | 7,072 | 3.6 | 1.6 | 1921 | 9,255 | 5.8 | 2.9 |
| 1917 | 7,269 | 3.8 | 1.8 | 1922 | 10,099 | 6.4 | 3.3 |
| 1918 | 7,484 | 4.0 | 1.9 | 1923 | 10,744 | 7.2 | 3.9 |
| 1919 | 7,788 | 4.3 | 2.1 | 1924 | 11,844 | 8.6 | 4.8 |
| 1920 | 8,624 | 5.0 | 2.5 | 1925 | 12,403 | 9.9 | 5.5 |

(a) Fit a straight line.
(b) Fit a parabola.

22. *p. 295.* Fire losses in the United States. Fit a straight line.

| *Year* | *Loss in Millions* | *Year* | *Loss in Millions* |
|---|---|---|---|
| 1916 | 258 | 1921 | 495 |
| 1917 | 290 | 1922 | 506 |
| 1918 | 354 | 1923 | 535 |
| 1919 | 320 | 1924 | 549 |
| 1920 | 448 | 1925 | 570 |

23. Personal Income Tax Returns are given in table XIV, appendix. Fit a curve of the type $xy = C$.

## Chapter XV

# THE NORMAL PROBABILITY CURVE AND THE PROBABLE ERROR

**167. Characteristics.**—A bell-shaped curve has been found to be the characteristic curve for many frequency distributions. The outstanding characteristics of this curve are:

(a) Symmetry with respect to the mean ordinate.
(b) Single mode, which is at the mean.
(c) Slope of the curve at the mean is zero.
(d) Slope of the curve approaches zero as $y$ approaches zero.

There are a number of curves, the equations of which are known, which satisfy these conditions:

I: $$y = Ae^{-Bx^2}$$

II: $$y = \frac{A}{B + Cx^2}$$

III: $$y = \frac{A}{e^x + e^{-x}}$$

It has been observed that the polygon determined by the terms in the expansion of

(1) $$N(\tfrac{1}{2} + \tfrac{1}{2})^n$$

has the first two properties listed.

G. Udny Yule has shown[1] that the frequency polygon given by the terms in the expansion of (1) approaches a definite limiting curve as $n$ increases without limit. The equation of this limiting curve is

$$y = y_o e^{-\frac{x^2}{2\sigma^2}}$$

[1] Yule, G. U. "An Introduction to the Theory of Statistics," Fourth Edition, Charles Griffin and Co. Limited, London, 1917, p. 301. In this derivation no calculus is used.

This is the form I given above and has the four characteristics (*a*), (*b*), (*c*), (*d*). It can be shown that

$$y_o = \frac{N}{\sigma\sqrt{2\pi}}$$

Hence, the equation[2] of the normal probability curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$$

**168. Normal probability curve.**—The equation of the normal probability curve may be put in the form

$$y = \frac{N}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} \tag{2}$$

In this equation $N$ = total frequency. $x$ is measured from the arithmetic mean. Whatever the original uniform class interval, it is convenient to call this interval unit. $\sigma$ = standard deviation, computed on the basis of a unit class interval.

In order to plot this curve, it is not necessary to compute from this equation values of $y$ for given values of $x$. Tables have been prepared which give the values of $Z$,

$$Z = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$$

for values of $u$ (table 87).

We can express the ordinates of (2) in terms of $Z$:

$$y = \frac{N}{\sigma}Z$$

[2] Kelley, Truman L., "Statistical Method," p. 94, Macmillan Co., gives a simple derivation involving the calculus.

See Rietz, H. L., "Handbook of Statistics," p. 96ff, for a derivation.

Briefly, the simplest differential equation which satisfies conditions (a), (b), (c), and (d) is $\frac{dy}{dx} = -k\,x\,y$. Separate the variables and integrate. If $y = y_o$ when $x = 0$, we find that

$$y = y_o e^{-\frac{k}{2}x^2}$$

See Forsyth, C. H., "Mathematical Analysis of Statistics" for a determination of $y_o$ and $k$.

If the class interval of the original frequency data is $C$ and not unity, and if the standard deviation $\sigma$ is expressed in terms of the units of the original data, then

$$y = \left(\frac{N}{\sigma}Z\right)C$$

To compute $y$ we proceed as follows:

1. Compute the arithmetic mean $Z$ of the original data.

2. In terms of unit class intervals, compute the deviations, $x$, from this mean.

3. On the basis of a unit class interval, compute the standard deviation $\sigma$.

4. Compute $u = \frac{x}{\sigma}\cdot$

5. Find from the table the values of $Z$ corresponding to the computed values of $u$.

6. Multiply the values of $Z$ by $\frac{N}{\sigma}\cdot$

Equation (2), in which $N$ is replaced by the total frequency and $\sigma$ is replaced by the value of the standard deviation as computed in step (3) above, is a normal probability curve which fits the given symmetrical frequency distribution.

The values of $y$ computed from (2) with given values of $N$ and $\sigma$ constitute a theoretical distribution with the same frequency and standard deviation as the given distribution. The values of $y$ so computed may be considered as a graduation or smoothing of the original data.

Let us illustrate with the following data[3] (table 86) which give the sums insured plus bonuses resulting from grouping a number of Endowment Insurances according to their office years of birth.

We try to obtain a knowledge of the characteristics of a population by the sampling process. Errors arise which are due to the sampling. In general, the larger the sample the smaller the errors. A graduation of the observed values of $y$, as outlined above, tends to eliminate the errors of

[3] Elderton, "Frequency Curves and Correlation," p. 88. Macmillan Co., N. Y.

sampling and gives us a more nearly accurate picture of the entire population from which the sample was obtained.

**Table 86.—Table Illustrating Graduation of Observed Values by Means of the Probability Curve.**

| *Central Age for 5 Groups of Years of Birth* $x'$ | *Sum Insured and Bonuses* ÷ 1,000 | *Class Interval Unity Origin at 112* $x''$ | *Deviations Class Interval Unity Origin 5.44* $x$ | $\frac{x}{\sigma}$ | $Z$ | *Graduation* $y = \frac{N}{\sigma}Z$ $= 753\ Z$ |
|---|---|---|---|---|---|---|
| 17 | 11 | 1 | −4.44 | 2.54 | .01585 | 12 |
| 22 | 48 | 2 | −3.44 | 1.97 | .05730 | 43 |
| 27 | 124 | 3 | −2.44 | 1.39 | .15183 | 115 |
| 32 | 213 | 4 | −1.44 | 0.82 | .28504 | 215 |
| 37 | 281 | 5 | −0.44 | 0.25 | .38667 | 291 |
| 42 | 295 | 6 | +0.56 | 0.32 | .37903 | 286 |
| 47 | 185 | 7 | 1.56 | 0.89 | .26848 | 202 |
| 52 | 104 | 8 | 2.56 | 1.46 | .13741 | 104 |
| 57 | 40 | 9 | 3.56 | 2.03 | .05082 | 38 |
| 62 | 15 | 10 | 4.56 | 2.61 | .01358 | 11 |
| 67 | 3 | 11 | 5.56 | 3.18 | .00262 | 2 |
| TOTAL...... | 1,319 = $N$ | | | | | |

$\sigma = 1.75$ (in terms of unit class intervals,

$\bar{x}' = 39.20$ years; $\bar{x}'' = 5.44$; $\frac{N}{\sigma} = \frac{1,319}{1.75} = 753.$

## Exercises.

1. The following table[4] gives the reserves resulting from grouping a number of Endowment Insurances according to their office years of birth. Graduate the reserves.

| *Central Age* | *Reserves ÷ 1,000* | *Central Age* | *Reserves ÷ 1,000* |
|---|---|---|---|
| 17 | 0.6 | 47 | 74.1 |
| 22 | 2.8 | 52 | 50.5 |
| 27 | 11.5 | 57 | 23.2 |
| 32 | 27.7 | 62 | 12.2 |
| 37 | 59.1 | 67 | 1.3 |
| 42 | 84.7 | | |
| | | | Total 347.7 |

Mean age = 43.97 $\qquad \frac{N}{\sigma} = \frac{347.7}{1.66} = 209.5$

$\sigma = 1.66$

[4] Elderton, "Frequency Curves and Correlation," p. 88. Macmillan Co., N. Y.

2. The following table[5] gives the frequency distribution of 750 students. Graduate the frequencies.

| *Height in Inches* | *No.* | *Height* | *No.* |
|---|---|---|---|
| 61 | 2 | 68 | 126 |
| 62 | 10 | 69 | 109 |
| 63 | 11 | 70 | 87 |
| 64 | 38 | 71 | 75 |
| 65 | 57 | 72 | 23 |
| 66 | 93 | 73 | 9 |
| 67 | 106 | 74 | 4 |
| | | | TOTAL 750 $= N$ |

Mean height $=$ 67.9 inches; $\sigma = 2.31$.

3. The following table gives the frequency distribution of weights of 993 ears of corn. Fit a normal frequency curve. $W =$ weight in ounces, $f =$ frequency.

| $W$ | $f$ | $W$ | $f$ | $W$ | $f$ | $W$ | $f$ |
|---|---|---|---|---|---|---|---|
| 2 | 4 | 8 | 75 | 14 | 112 | 20 | 2 |
| 3 | 22 | 9 | 71 | 15 | 65 | 21 | 1 |
| 4 | 27 | 10 | 75 | 16 | 37 | | |
| 5 | 50 | 11 | 88 | 17 | 8 | TOTAL | 993 |
| 6 | 47 | 12 | 107 | 18 | 13 | | |
| 7 | 71 | 13 | 114 | 19 | 4 | | |

Mean weight $=$ 10.65; $\sigma = 3.63$.

4. The following table gives the frequency distribution of length of heads for Cairo-born Egyptians. Fit a normal frequency curve and graduate the frequencies. $L =$ length in mm.; $F =$ observed frequency; $G =$ graduated frequency. [*Biometrika*, Vol. XI, p. 72].

$$y = 202.9\, e^{-\frac{x^2}{69.6}}, \text{ where } x \text{ is measured from the mean.}$$

| $L$ | $F$ | $G$ | $L$ | $F$ | $G$ | $L$ | $F$ | $G$ |
|---|---|---|---|---|---|---|---|---|
| 174 | 4 | 3 | 186 | 114 | 122 | 198 | 75 | 73 |
| 177 | 8 | 12 | 189 | 177 | 158 | 201 | 39 | 34 |
| 180 | 41 | 33 | 192 | 160 | 158 | 204 | 7 | 12 |
| 183 | 69 | 72 | 195 | 98 | 122 | 207 | 10 | 3 |

[5] West, C. J. "Introduction to Mathematical Statistics," p. 46, p. 50. Columbus, R. G. Adams, 1918.

Table 87.—Ordinates of Probability Curves.

$$u = \frac{x}{\sigma}$$

| $u$ | $Z$ | $u$ | $Z$ | $u$ | $Z$ | $u$ | $Z$ |
|---|---|---|---|---|---|---|---|
| .00 | .39894 | 1.00 | .24197 | 2.00 | .05399 | 3.00 | .00443 |
| .05 | .39844 | 1.05 | .22988 | 2.05 | .04879 | 3.05 | .00381 |
| .10 | .39695 | 1.10 | .21785 | 2.10 | .04398 | 3.10 | .00327 |
| .15 | .39448 | 1.15 | .20594 | 2.15 | .03955 | 3.15 | .00279 |
| .20 | .39104 | 1.20 | .19419 | 2.20 | .03547 | 3.20 | .00238 |
| .25 | .38667 | 1.25 | .18265 | 2.25 | .03174 | 3.25 | .00203 |
| .30 | .38139 | 1.30 | .17137 | 2.30 | .02833 | 3.30 | .00173 |
| .35 | .37524 | 1.35 | .16038 | 2.35 | .02522 | 3.35 | .00146 |
| .40 | .36827 | 1.40 | .14973 | 2.40 | .02239 | 3.40 | .00123 |
| .45 | .36053 | 1.45 | .13943 | 2.45 | .01984 | 3.45 | .00104 |
| .50 | .35207 | 1.50 | .12952 | 2.50 | .01753 | 3.50 | .00087 |
| .55 | .24294 | 1.55 | .12001 | 2.55 | .01545 | 3.55 | .00073 |
| .60 | .33322 | 1.60 | .11092 | 2.60 | .01358 | 3.60 | .00061 |
| .65 | .32297 | 1.65 | .10226 | 2.65 | .01191 | 3.65 | .00051 |
| .70 | .31225 | 1.70 | .09405 | 2.70 | .01042 | 3.70 | .00042 |
| .75 | .30114 | 1.75 | .08628 | 2.75 | .00909 | 3.75 | .00035 |
| .80 | .28969 | 1.80 | .07895 | 2.80 | .00792 | 3.80 | .00029 |
| .85 | .27798 | 1.85 | .07206 | 2.85 | .00687 | 3.85 | .00024 |
| .90 | .26609 | 1.90 | .06562 | 2.90 | .00595 | 3.90 | .00020 |
| .95 | .25406 | 1.95 | .05959 | 2.95 | .00514 | 3.95 | .00016 |
| | | | | | | 4.00 | .00013 |

*Adapted from Karl Pearson, "Tables for Statisticians and Biometricians."*

**169. Probable error.**—The items of a frequency distribution have different measures. Compute the arithmetic average of these measures. Compute for each item the deviation of its measure from the computed arithmetic average. Give positive signs to all of these deviations, and arrange them in order of magnitude. The *median* of this list is called the *probable error* (P.E.) of the set of observations. It can be shown that

$$\text{P.E.} = 0.6745\sigma$$

In a normal distribution, if an item is selected at random, it is an even chance that the deviation of its measure from the arithmetic average will not be more than the probable error.

The probable error, for a normal distribution, must agree with the semi-interquartile range, for the measures of exactly half of the items fall between the first quartile ($Q_1$) and the third quartile ($Q_3$) and $\overline{X} = Q_2 = \frac{1}{2}(Q_1 + Q_3)$.

For a binomial distribution $\sigma = \sqrt{npq}$ and hence for a binomial distribution

$$\text{P.E.} = 0.6745\sqrt{npq}$$

**170. Probable error in the arithmetic average.**[6]—Take a sample of 500 ears of corn from a crib. Compute the arithmetic average of their lengths. We use this to represent the average length of all the ears in the crib. Quite probably it differs from their true arithmetic average. We now find, by means of equation (3) below, a number P.E.$_{\overline{X}}$

$$\text{(3)} \qquad \text{P.E.}_X = 0.6745\frac{\sigma}{\sqrt{N}}$$

called the *probable error in the arithmetic average.* This is a number such that it is equally likely whether or not the computed arithmetic average of the 500 ears selected lies between $X$ − P.E.$_{\overline{X}}$ and $\overline{X}$ + P.E.$_{\overline{X}}$, where $\overline{X}$ denotes the true (unknown) arithmetic average for all the ears in the crib. In other words, if a very large number of persons take a sample of ears and each computes an average length, then, in a sufficiently large number of cases, one-half of these averages will be within the limits set and one-half will be without.

In treatises on probability it is shown that equation (3) above is true. This formula shows that in order to double the precision of the computed arithmetic average, it is necessary to take four times as many observations.

For example, let us determine the average weekly wage of 20,000 coal miners by taking a sample consisting of 256 workers. Suppose that an average computed from this sample is $40.00 with a standard deviation of $2.40. What

[6] Adapted from Kenyon-Lovitt, "Mathematics for Students of Agriculture and General Science," p. 300, Macmillan Co.

is the reliability of these results? The probable error of the arithmetic average is

$$\text{P.E.}_{\bar{X}} = 0.6745\frac{2.40}{\sqrt{256}} = 0.101$$

This means that if another sample of 256 was to be taken, the chances are even that the arithmetic average computed from the new sample would be within the limits $40.00 ± 0.101. Expressed differently, this means that if many samples of the same size were to be taken, and an arithmetic average computed for each, it is to be expected that one-half of these averages will fall within the limits computed, namely $39.90 and $40.10.

Suppose another sample of 1,024 workers to have the same arithmetic average and standard deviation. The probable error in the arithmetic average would be smaller, that is, only half as large as for the sample of 256 workers:

$$\text{P.E.}_{\bar{X}} = 0.6745\frac{\sqrt{2.40}}{1{,}024} = 0.05$$

**171. Probable error in the standard deviation.**—Compute the standard deviation for the weekly wage of a sample of 512 coal miners. This will differ slightly from the true standard deviation $\sigma$, of the weekly wage of the group from which the sample was taken. Next find by means of the formula (4) the probable error (P.E.$_\sigma$) of the standard deviation. Then

$$\text{P.E.}_\sigma = 0.6745\frac{\sigma}{\sqrt{2N}} \tag{4}$$

For a sufficiently large number of samples from the entire group, the computed standard deviations will fall one-half within the limits $\sigma -$ P.E.$_\sigma$ and $\sigma +$ P.E.$_\sigma$ and one-half without.

If the standard deviation is $2.40, then

$$\text{P.E.}_\sigma = 0.6745\frac{2.40}{\sqrt{2 \times 512}} = 0.05$$

We observe from formula (4) that it is necessary to increase the size of the sample fourfold in order to halve the probable error.

**172. Probable error of the relative frequency.**[7]—We have seen that the probable error in a binomial distribution is $0.6745\sqrt{npq}$ where $n$ is the size of the sample and $p$ is the probability that an event will happen and $q$ the probability that the event will not happen.

To illustrate: suppose a community of 2,500 to consist of 900 blacks and 1,600 whites. In any representative sample of 100, the number of blacks should, theoretically, be 36. The chances are even that the number of blacks actually included in the sample of 100 will differ from 36 by not more than

$$0.6745\sqrt{100 \times \tfrac{9}{25} \times \tfrac{16}{25}} = 3.217$$

for $n = 100$, $p = \frac{9}{25}$, $q = \frac{16}{25}$. This number, 3.217, is called the *probable error of the frequency.*

The ratio of this number to the total number ($n$) in the sample is called the *probable error of the relative frequency.* In this case $n = 100$, hence

$$\text{P.E. of relative frequency} = 0.03217$$

In general

$$\text{P.E. of relative frequency} = \frac{0.6745\sqrt{npq}}{n}$$
$$= 0.6745\sqrt{\frac{pq}{n}}$$

If $f$ is the number of successes in $n$ trials, then $p = \dfrac{f}{n}$ $q = 1 - \dfrac{f}{n}$, and

$$\text{P.E. of relative frequency} = 0.6745\sqrt{\frac{\frac{f}{n}\left(1 - \frac{f}{n}\right)}{n}}$$

In practice, we do not know the value of $p$. We endeavor to find an approximate value of $p$ from the items in a sample. For $p$ we take the proportion found in the sample.

For example, let us endeavor to find the proportion of the population in a given district infected with hookworm.

[7] See Rietz, H. L., "Handbook of Mathematical Statistics," p. 74, Houghton Mifflin Co., Boston, 1924.

Suppose we take, to the best of our ability, a representative sample of 100 individuals. Suppose we find 36 infected. Then the proportion $p$ found from the sample is $p = \frac{36}{100} = \frac{9}{25}$. Hence $q = \frac{16}{25}$. Computing, we find the probable error of relative frequency to be

$$0.6745\sqrt{\frac{\frac{9}{25}\frac{16}{25}}{100}} = 0.03217$$

That is, the chances are even that the percentage of persons in the district under consideration who are infected does not differ from 36 per cent by more than 3.2 per cent. That is, it is an even chance that the value of $p$ is between

$$0.36 - 0.032 = 0.328 \text{ and } 0.36 + 0.032 = 0.392$$

The ratio of any measure ($m$) to its probable error ($e$) is some indication of the reliability to be attached to the measure. The odds against the occurrence of a ratio greater than $\frac{m}{e}$ is given in table 88. From this table we see that the greater $\frac{m}{e}$, the greater the reliability.

**Table 88.—Form Showing the Relation of Reliability to $\frac{m}{e}$.**

| $\frac{m}{e} = z$ | *Odds against the Occurrence of a Ratio Greater than Z.* |
|---|---|
| 1 | 1 to 1 |
| 2 | 4.6 to 1 |
| 3 | 22 to 1 |
| 4 | 142 to 1 |
| 5 | 1,300 to 1 |
| 6 | 20,000 to 1 |

**173. Probable error in the difference of two measures.**—Let $m_1$ and $m_2$ be the measures under consideration. These measures may be arithmetic average, standard deviation, or any other. Suppose that their respective probable errors are P.E.$_1$ and P.E.$_2$. It can be shown that the probable error of the difference $m_1 - m_2$ is

$$\sqrt{(\text{P.E.}_1)^2 + (\text{P.E.}_2)^2}$$

Professor Pearl[8] gives an illustration of the use of this formula. On examination of 150 people, the average number of pulse beats per minute was found to be 79.68 with a probable error of 0.15. This is usually written 79.68 ± 0.15. After the administration of a certain drug the average number of pulse beats per minute was found to be 81.12 ± 0.20. Is this increase significant? That is, is it highly probable that the increase is due to the drug, or is the increase a result of chance due to the sample? We have

$$m_1 - m_2 = 81.12 - 79.68 = 1.44$$
$$\text{P.E. of } (m_1 - m_2) = \sqrt{(0.15)^2 + (0.20)^2 = 0.25}$$

Whence

$$\frac{1.44}{0.25} = 5.96$$

From table 88, we see that the chances are nearly 20,000 to 1 that the change in pulse beat is due to the drug.

This formula should be useful in comparing the costs of two different kinds of school administration, factory management, methods of production, methods of treating a disease.

**174. Probable errors of various statistical constants** have been computed. For handy reference, the formulas for a few of these are here given.[9] $N$ is the number of observations.

$$\text{P.E. of any distribution} = 0.6745\sigma$$
$$\text{P.E. of } \overline{X} \text{ (arithmetic average)} = 0.6745\frac{\sigma}{\sqrt{N}}$$
$$\text{P.E. of } \sigma \text{ (standard deviation)} = 0.6745\frac{\sigma}{\sqrt{2N}}$$
$$\text{P.E. of median} = 0.8454\frac{\sigma}{\sqrt{N}}$$
$$\text{P.E. of relative frequency} = 0.6745\sqrt{\frac{pq}{N}}$$
$$\text{P.E. of } r \text{ (coefficient of correlation)} = 0.6745\ \frac{1 - r^2}{\sqrt{N}}$$

[8] Pearl, Raymond, "Medical Biometry and Statistics," p. 214. Saunders, Philadelphia, 1923.

[9] Rietz, H. L., "Handbook," loc. cit. p. 77.

**Exercises.**

1. *p. 138.* (*Source: Indices of General Business Conditions, W. M. Persons, Cambridge, Mass., 1919.*) Fit a normal curve to the following data:

Irregular fluctuations of building permits issued for twenty leading cities. Frequency table by months, July, 1903–June, 1916.

| *Irregular Fluctuations, Percentages* | *Frequency* | *Irregular Fluctuations, Percentages* | *Frequency* |
|---|---|---|---|
| +46.0 | 1 | + 1.9 | 23 |
| +35.0 | 1 | − 2.1 | 19 |
| +32.3 | 1 | − 6.1 | 19 |
| +21.9 | 4 | −10.1 | 15 |
| +17.9 | 8 | −14.1 | 9 |
| +13.9 | 10 | −18.1 | 4 |
| + 9.9 | 15 | −22.1 | 3 |
| + 5.9 | 23 | −26.1 | 1 |
| | | | TOTAL 156 |

*Ans.* $y = 21.7e^{-\frac{x^2}{16.48}}$

2. Compute the probable error:

(a) of the standard deviation for the data table 45;

(b) of the coefficient of correlation for the data table 51; and for the exercises at the end of §90;

(c) of the arithmetic average and standard deviation for the data in exercise 10, §149.

# APPENDIX

# APPENDIX

**Table I.—Frequency Distribution of Amounts Paid for Meals at McRae's Restaurant, Colorado Springs, Jan. 21, 22, 1929 (A few amounts over 1.25 have been omitted).**

| | Breakfast | | Dinner | | Supper | | 8 P.M. to 2 A.M. | |
|---|---|---|---|---|---|---|---|---|
| *Price* | *Jan. 21* | *Jan. 22* | *Jan. 21* | *Jan. 22* | *Jan. 21* | *Jan. 22* | *Jan. 21* | *Jan. 22* |
| 0.05 | 11 | 21 | 16 | 11 | 12 | 4 | 8 | 6 |
| 0.10 | 7 | 7 | 11 | 6 | 9 | 7 | 2 | 6 |
| 0.15 | 15 | 21 | 18 | 18 | 15 | 16 | 5 | 2 |
| 0.20 | 24 | 14 | 7 | 9 | 10 | 7 | 4 | 3 |
| 0.25 | 25 | 29 | 11 | 12 | 7 | 4 | 0 | 1 |
| 0.30 | 12 | 13 | 11 | 12 | 8 | 11 | 1 | 5 |
| 0.35 | 21 | 18 | 7 | 10 | 5 | 4 | 1 | 1 |
| 0.40 | 15 | 9 | 9 | 7 | 2 | 3 | 3 | 1 |
| 0.45 | 11 | 19 | 5 | 10 | 6 | 3 | 0 | 4 |
| 0.50 | 11 | 11 | 2 | 6 | 1 | 2 | 5 | 4 |
| 0.55 | 8 | 12 | 0 | 2 | 1 | 1 | 8 | 4 |
| 0.60 | 8 | 7 | 3 | 11 | 4 | 1 | 5 | 8 |
| 0.65 | 5 | 6 | 0 | 4 | 3 | 2 | 5 | 7 |
| 0.70 | 5 | 2 | 1 | 1 | 0 | 1 | 1 | 5 |
| 0.75 | 3 | 4 | 2 | 0 | 1 | 1 | 3 | 7 |
| 0.80 | 1 | 6 | 2 | 2 | 1 | 1 | 1 | 5 |
| 0.85 | 1 | 3 | 1 | 2 | 3 | 1 | 8 | 4 |
| 0.90 | 1 | 2 | 2 | 1 | 0 | 0 | 1 | 2 |
| 0.95 | 0 | 2 | .. | 0 | 0 | 3 | 2 | 2 |
| 1.00 | 1 | 0 | .. | 1 | 0 | 2 | 3 | 4 |
| 1.05 | 1 | 3 | .. | .. | 1 | 0 | 0 | 0 |
| 1.10 | 2 | 2 | .. | .. | 0 | 0 | 1 | 0 |
| 1.15 | 0 | 0 | .. | .. | 1 | 1 | 0 | 0 |
| 1.20 | 0 | 0 | .. | .. | .. | .. | 1 | 0 |
| 1.25 | 0 | 1 | .. | .. | .. | .. | 0 | 2 |

**Table II.—Frequency Distribution in First-Year Egg Production.**

| *Annual Egg Production* | 0 | 15 | 30 | 45 | 60 | 75 | 90 | 105 | 120 | 135 | 150 | 165 | 180 | 195 | 210 | 225 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 14 | 29 | 44 | 59 | 74 | 89 | 104 | 119 | 134 | 149 | 164 | 179 | 194 | 209 | 224 | 239 |
| 1902–03 | . . . | 2 | 0 | 1 | 5 | 8 | 17 | 18 | 17 | 26 | 17 | 18 | 9 | 2 | 6 | 1 |
| 1903–04 | 7 | 5 | 5 | 10 | 10 | 20 | 24 | 29 | 52 | 37 | 29 | 16 | 8 | 2 | | |
| 1905–06 | . . . | . . . | 1 | 2 | 4 | 9 | 13 | 25 | 24 | 22 | 32 | 17 | 20 | 9 | | |

1902–03: $\overline{X} = 135.98$; S.D. = 39.5; Mo = 142.5; Md = 138.17
1903–04: $\overline{X} = 117.37$; S.D. = 41.62; Mo = 128.41; Md = 124.9
1905–06: $\overline{X} = 139.8$; S.D. = 36.02; Mo = 156.5; Md = 142.5

*Bulletin 110, Part I, Bureau of Animal Husbandry, U. S. Dep't. of Agriculture, "A Biometrical Study of Egg Production in the Domestic Fowl."*

**Table III.—Distribution of Grades of 127 Colorado College Freshmen in English for the Spring of 1923, as Given by the Registrar's Records.**

| | | | |
|---|---|---|---|
| 50 and under 60 | 2 | 80 and under 90 | 40 |
| 60 and under 70 | 18 | 90 and under 100 | 23 |
| 70 and under 80 | 44 | | |

**Table IV.—Estimated Distribution of Income among the Single Women of the Continental United States in 1910.**

| *Income $* | *Number* | *Income $* | *Number* |
|---|---|---|---|
| 0–200 | 10 | 800– 900 | 37 |
| 200–300 | 70 | 900–1,000 | 22 |
| 300–400 | 560 | 1,000–1,100 | 16 |
| 400–500 | 530 | 1,100–1,200 | 12 |
| 500–600 | 280 | 1,200–1,300 | 8 |
| 600–700 | 150 | 1,300–1,400 | 5 |
| 700–800 | 110 | | |

*King, "Wealth and Income," p. 224.*

Table V.—Earnings of Male Employees in the United States.

| Earnings per Hour | Number Receiving the Earnings Stated | |
|---|---|---|
| | *Chemicals* | *Lumber* |
| Total | 28,478 | 18,022 |
| Under 20¢ | 75 | 617 |
| 20 and under 30¢ | 1,217 | 5,118 |
| 30 40 | 7,903 | 6,398 |
| 40 50 | 10,525 | 2,702 |
| 50 60 | 4,470 | 2,619 |
| 60 70 | 2,520 | 370 |
| 70 80 | 1,215 | 64 |
| 80 90 | 448 | 75 |
| 90 $1.00 | 63 | 18 |
| 1.00 1.25 | 24 | 33 |
| 1.25 1.50 | 18 | 8 |
| Average earnings | $0.466 | $0.369 |

*U. S. Bureau of Labor Statistics, Monthly Labor Review, Sept., 1919.*

**Table VI.—Roving-Frame Tenders in Cotton Mills: Females 16 Years of Age and Over.**

| *Rates per Week* DOLLARS | FREQUENCIES | |
|---|---|---|
| | *1890* | *1900* |
| 2.50–2.99 | 2 | 0 |
| 3.00–3.49 | 8 | 1 |
| 3.50–3.99 | 13 | 0 |
| 4.00–4.49 | 9 | 8 |
| 4.50–4.99 | 34 | 17 |
| 5.00–5.49 | 50 | 31 |
| 5.50–5.99 | 52 | 36 |
| 6.00–6.49 | 103 | 66 |
| 6.50–6.99 | 68 | 61 |
| 7.00–7.49 | 31 | 53 |
| 7.50–7.99 | 11 | 25 |
| 8.00–8.49 | 17 | 42 |
| 8.50–8.99 | 1 | 39 |
| 9.00–9.49 | 1 | 24 |
| 9.50–9.99 | 0 | 6 |

$$1890 \begin{cases} \overline{X} = 6.03 \\ \text{Mo} = 6.28 \\ \text{Md} = 6.155 \end{cases}$$

$$1900 \begin{cases} \overline{X} = 6.99 \\ \text{Mo} = 6.31 \\ \text{Md} = 6.87 \end{cases}$$

*Special Report on Employees and Wages, U. S. Census, 1900, p. 31.*

**Table VII.**

| | *Wheat* | *Corn* | *Oats* | *Barley* | *Rye* | *Flax Seed* |
|---|---|---|---|---|---|---|
| *Production 1,000 bu.....* | 785,741 | 3,054,395 | 1,299,823 | 198,185 | 63,023 | 17,429 |
| Average farm price per bu. Dec. 1........... | 92.3 | 72.7 | 41.5 | 54.0 | 64.7 | 210.8 |

*Yearbook of Dep't of Agriculture, 1923.*

Table VIII.

| | *Nebr.* | *Kans.* | *Iowa* | *Okla.* | *Wis.* | *Minn.* |
|---|---|---|---|---|---|---|
| Wheat production 1,000 bu., 1923. | 31,388 | 83,804 | 14,352 | 36,300 | 1,970 | 20,785 |
| Average farm price per bu....... | 83.0 | 91.0 | 89.0 | 93.0 | 98.0 | 95 |

*Yearbook of Dep't of Agriculture, 1923*

Table IX.

| | *Corn* | *Wheat* | *Oats* | *Barley* | *Rye* | *Potatoes* |
|---|---|---|---|---|---|---|
| Price, cents per bu........... | 62.51 | 95.33 | 37.58 | 69.41 | 63.62 | 59.32 |
| Production, million bu........ | 2,447 | 763.4 | 1,122 | 178.2 | 41.38 | 331.5 |

*U. S. Bureau of Labor Statistics, Bull. 181; Yearbook, Dept. of Agriculture, 1923.*

Table X.

| *Commodity* | *Average Quantity per Family in the U. S., Lbs.* | *Cost per Lb. in ¢* | *Commodity* | *Quantity* | *Cost* |
|---|---|---|---|---|---|
| RICE......... | 25.1 | 8.16 | BUTTER........ | 117.1 | 24.56 |
| SUGAR....... | 268.5 | 5.87 | FLOUR AND MEAL | 680.8 | 2.46 |
| COFEE....... | 46.8 | 22.95 | FRESH BEEF..... | 349.7 | 14.31 |
| TEA......... | 10.6 | 50.00 | | | |

*Eighteenth Annual Report of the Commissioner of Labor, 1903, p. 648.*

### Table XI.

| | North Atlantic States<br>1,415 *Families* | | North Central States<br>721 *Families* | |
|---|---|---|---|---|
| *Commodity* | *Lbs.* | *Cents per Lb.* | *Lbs.* | *Cents per Lb.* |
| Rice | 22.2 | 8.74 | 21.8 | 8.81 |
| Sugar | 282.8 | 5.89 | 253.1 | 5.77 |
| Coffee | 38.5 | 25.19 | 57.5 | 22.28 |
| Tea | 12.9 | 49.07 | 8.5 | 49.65 |
| Butter | 118.9 | 25.04 | 124.0 | 22.97 |
| Flour and meal | 624.0 | 2.60 | 718.2 | 2.29 |
| Fresh beef | 352.2 | 15.41 | 363.5 | 12.67 |

*Eighteenth Annual Report of Commissioner of Labor, 1903, p. 648.*

### Table XII.—American Experience Mortality Table. Based on 100,000 Living at Age 10.

| *At Age* | *Number Surviving* | *Deaths* | *At Age* | *Number Surviving* | *Deaths* | *At Age* | *Number Surviving* | *Deaths* | *At Age* | *Number Surviving* | *Deaths* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 100,000 | 749 | 32 | 84,000 | 723 | 54 | 65,706 | 1,143 | 76 | 23,761 | 2,431 |
| 11 | 99,251 | 746 | 33 | 83,277 | 726 | 55 | 64,563 | 1,199 | 77 | 21,330 | 2,369 |
| 12 | 98,505 | 743 | 34 | 82,551 | 729 | 56 | 63,364 | 1,260 | 78 | 18,961 | 2,291 |
| 13 | 97,762 | 740 | 35 | 81,822 | 732 | 57 | 62,104 | 1,325 | 79 | 16,670 | 2,196 |
| 14 | 97,022 | 737 | 36 | 81,090 | 737 | 58 | 60,779 | 1,394 | 80 | 14,474 | 2,091 |
| 15 | 96,285 | 735 | 37 | 80,353 | 742 | 59 | 59,385 | 1,468 | 81 | 12,383 | 1,964 |
| 16 | 95,550 | 732 | 38 | 79,611 | 749 | 60 | 57,917 | 1,546 | 82 | 10,419 | 1,816 |
| 17 | 94,818 | 729 | 39 | 78,862 | 756 | 61 | 56,371 | 1,628 | 83 | 8,603 | 1,648 |
| 18 | 94,089 | 727 | 40 | 78,106 | 765 | 62 | 54,743 | 1,713 | 84 | 6,955 | 1,470 |
| 19 | 93,362 | 725 | 41 | 77,341 | 774 | 63 | 53,030 | 1,800 | 85 | 5,485 | 1,292 |
| 20 | 92,637 | 723 | 42 | 76,567 | 785 | 64 | 51,230 | 1,889 | 86 | 4,193 | 1,114 |
| 21 | 91,914 | 722 | 43 | 75,782 | 797 | 65 | 49,341 | 1,980 | 87 | 3,079 | 933 |
| 22 | 91,192 | 721 | 44 | 74,985 | 812 | 66 | 47,361 | 2,070 | 88 | 2,146 | 744 |
| 23 | 90,471 | 720 | 45 | 74,173 | 828 | 67 | 45,291 | 2,158 | 89 | 1,402 | 555 |
| 24 | 89,751 | 719 | 46 | 73,345 | 848 | 68 | 43,133 | 2,243 | 90 | 847 | 385 |
| 25 | 89,032 | 718 | 47 | 72,497 | 870 | 69 | 40,890 | 2,321 | 91 | 462 | 246 |
| 26 | 88,314 | 718 | 48 | 71,627 | 896 | 70 | 38,569 | 2,391 | 92 | 216 | 137 |
| 27 | 87,596 | 718 | 49 | 70,731 | 927 | 71 | 36,178 | 2,448 | 93 | 79 | 58 |
| 28 | 86,878 | 718 | 50 | 69,804 | 962 | 72 | 33,730 | 2,487 | 94 | 21 | 18 |
| 29 | 86,160 | 719 | 51 | 68,842 | 1,001 | 73 | 31,243 | 2,505 | 95 | 3 | 3 |
| 30 | 85,441 | 720 | 52 | 67,841 | 1,044 | 74 | 28,738 | 2,501 | | | |
| 31 | 84,721 | 721 | 53 | 66,797 | 1,091 | 75 | 26,237 | 2,476 | | | |

Table XIII.—Distribution of the Relative Prices of 1,437 Different Commodities in 1918. (Average Prices in July, 1913, to June, 1914 = 100.)

| *Relative Price* | *Number of Cases* | *Relative Price* | *Number of Cases* | *Relative Price* | *Number of Cases* | *Relative Price* | *Number of Cases* |
|---|---|---|---|---|---|---|---|
| 36 | 1 | 250–269 | 76 | 490–509 | 4 | 848 | 1 |
| 49 | 1 | 270–289 | 54 | 510–529 | 5 | 900 | 1 |
| 50– 69 | 4 | 290–309 | 42 | 530–549 | 3 | 1,165 | 1 |
| 70– 89 | 17 | 310–329 | 30 | 550–569 | 4 | 1,356 | 1 |
| 90–109 | 61 | 330–349 | 31 | 587 | 1 | 1,585 | 1 |
| 110–129 | 64 | 350–369 | 16 | 627 | 1 | 1,764 | 1 |
| 130–149 | 130 | 370–389 | 13 | 727 | 1 | 2,049 | 1 |
| 150–169 | 212 | 390–409 | 7 | 730 | 1 | 2,863 | 1 |
| 170–189 | 219 | 410–429 | 7 | 743 | 1 | 3,009 | 1 |
| 190–209 | 164 | 430–449 | 8 | 761 | 1 | | |
| 210–229 | 135 | 450–469 | 4 | 784 | 1 | | |
| 230–249 | 104 | 470–489 | 4 | 826 | 1 | | |

$\overline{X} = 219.6$; $G = 198.3$; Md = 191.2; Mo. = 174.4.

*Bulletin No. 284 of the U. S. Bureau of Labor Statistics, p. 70.*

Table XIV.—Personal Income Tax Returns.

| *Income Class* | *Number of Returns* | *Income Class* | *Number of Returns* |
|---|---|---|---|
| Under $1,000 | 402,076 | 25,000– 50,000 | 35,478 |
| 1,000– 2,000 | 2,471,181 | 50,000– 100,000 | 12,000 |
| 2,000– 3,000 | 2,129,898 | 100,000– 150,000 | 2,171 |
| 3,000– 4,000 | 588,065 | 150,000– 300,000 | 1,323 |
| 4,000– 5,000 | 380,594 | 300,000– 500,000 | 309 |
| 5,000– 6,000 | 383,189 | 500,000–1,000,000 | 161 |
| 6,000–10,000 | 237,184 | 1,000,000 and over | 67 |
| 10,000–25,000 | 151,329 | | |

*Statistical Abstract of U. S., 1924, p. 163.*

**Table XV.—Average Farm Price, Dec. 1.**

| *Year* | *Corn* | *Wheat* | *Oats* | *Barley* | *Rye* | *Potatoes* | *Hay $ Ton* |
|---|---|---|---|---|---|---|---|
| *1870* | 49.4 | 94.4 | 39.0 | 79.1 | 73.2 | 65.0 | 12.47 |
| *71* | 43.4 | 114.5 | 36.2 | 75.8 | 71.1 | 53.9 | 14.30 |
| *72* | 35.3 | 111.4 | 29.9 | 68.6 | 67.6 | 53.5 | 12.94 |
| *73* | 44.2 | 106.9 | 34.6 | 86.7 | 70.3 | 65.2 | 12.53 |
| *74* | 58.4 | 86.3 | 47.1 | 86.0 | 77.4 | 61.5 | 11.94 |
| *1875* | 36.7 | 89.5 | 32.0 | 74.1 | 67.1 | 34.4 | 10.78 |
| *76* | 34.0 | 97.0 | 32.4 | 63.0 | 61.4 | 61.9 | 8.97 |
| *77* | 34.8 | 105.7 | 28.4 | 62.5 | 57.6 | 43.7 | 8.37 |
| *78* | 31.7 | 77.6 | 24.6 | 57.9 | 52.5 | 58.7 | 7.20 |
| *79* | 37.5 | 110.8 | 33.1 | 58.9 | 67.6 | 43.6 | 9.32 |
| *1880* | 39.6 | 95.1 | 36.0 | 66.6 | 75.6 | 48.3 | 11.65 |
| *81* | 63.6 | 119.2 | 46.4 | 82.3 | 93.3 | 91.0 | 11.82 |
| *82* | 48.5 | 88.4 | 37.5 | 62.9 | 61.5 | 55.7 | 9.73 |
| *83* | 42.4 | 91.1 | 32.7 | 58.7 | 58.1 | 42.2 | 8.19 |
| *84* | 35.7 | 64.5 | 27.7 | 48.7 | 51.9 | 39.6 | 8.17 |
| *1885* | 32.8 | 77.1 | 28.5 | 56.3 | 57.9 | 44.7 | 8.71 |
| *86* | 36.6 | 68.7 | 29.8 | 53.6 | 53.8 | 46.7 | 8.46 |
| *87* | 44.4 | 68.1 | 30.4 | 51.9 | 54.5 | 68.2 | 9.97 |
| *88* | 34.1 | 92.6 | 27.8 | 59.0 | 58.8 | 40.2 | 8.76 |
| *89* | 28.3 | 69.8 | 22.9 | 41.6 | 42.3 | 35.4 | 7.04 |
| *1890* | 50.6 | 83.8 | 42.4 | 62.7 | 62.9 | 75.8 | 7.87 |
| *91* | 40.6 | 83.9 | 31.5 | 52.4 | 77.4 | 35.8 | 8.12 |
| *92* | 39.4 | 62.4 | 31.7 | 47.5 | 54.2 | 66.1 | 8.20 |
| *93* | 36.5 | 53.8 | 29.4 | 41.1 | 51.3 | 59.4 | 8.68 |
| *94* | 45.7 | 49.1 | 32.4 | 44.2 | 50.1 | 53.6 | 8.54 |
| *1895* | 25.3 | 50.9 | 19.9 | 33.7 | 44.0 | 26.6 | 8.35 |
| *96* | 21.5 | 72.6 | 18.7 | 32.3 | 40.9 | 28.6 | 6.55 |
| *97* | 26.3 | 80.8 | 21.2 | 37.7 | 44.7 | 54.7 | 6.62 |
| *98* | 28.7 | 58.2 | 25.5 | 41.3 | 46.3 | 41.4 | 6.00 |
| *99* | 30.3 | 58.4 | 24.9 | 40.3 | 51.0 | 39.0 | 7.27 |
| *1900* | 35.7 | 61.9 | 25.8 | 40.9 | 51.2 | 43.1 | 8.89 |
| *01* | 60.5 | 62.4 | 39.9 | 45.2 | 55.7 | 76.7 | 10.01 |
| *02* | 40.3 | 63.0 | 30.7 | 45.9 | 50.8 | 47.1 | 9.06 |
| *03* | 42.5 | 69.5 | 34.1 | 45.6 | 54.5 | 61.4 | 9.07 |
| *04* | 44.1 | 92.4 | 31.3 | 42.0 | 68.8 | 45.3 | 8.72 |
| *1905* | 41.2 | 74.8 | 29.1 | 40.5 | 61.1 | 61.7 | 8.52 |
| *06* | 39.9 | 66.7 | 31.7 | 41.5 | 58.9 | 51.1 | 10.37 |
| *07* | 51.6 | 87.4 | 44.3 | 66.6 | 73.1 | 61.8 | 11.68 |
| *08* | 60.6 | 92.8 | 47.2 | 55.4 | 73.6 | 70.6 | 8.98 |
| *09* | 57.9 | 98.6 | 40.2 | 54.0 | 71.8 | 54.1 | 10.50 |
| *1910* | 48.0 | 88.3 | 34.4 | 57.8 | 71.5 | 55.7 | 12.14 |
| *11* | 61.8 | 87.4 | 45.0 | 86.9 | 83.2 | 79.9 | 14.29 |
| *12* | 48.7 | 76.0 | 31.9 | 50.5 | 66.3 | 50.5 | 11.79 |
| *13* | 69.1 | 79.9 | 39.2 | 53.7 | 63.4 | 68.7 | 12.43 |
| *14* | 64.4 | 98.6 | 43.8 | 54.3 | 86.5 | 48.9 | 11.12 |
| *1915* | 57.5 | 92.0 | 36.1 | 51.6 | 83.4 | 61.6 | 10.63 |
| *16* | 88.9 | 160.3 | 52.4 | 88.1 | 122.1 | 146.1 | 11.22 |
| *17* | 127.9 | 200.8 | 66.6 | 113.7 | 166.0 | 122.8 | 17.09 |
| *18* | 136.5 | 204.2 | 70.9 | 91.7 | 151.6 | 119.3 | 20.13 |
| *19* | 134.5 | 214.9 | 70.4 | 120.6 | 133.2 | 159.5 | 20.05 |
| *1920* | 67.0 | 143.7 | 46.0 | 71.3 | 126.8 | 114.5 | 17.66 |
| *21* | 42.3 | 92.6 | 30.2 | 41.9 | 69.7 | 110.1 | 12.10 |
| *22* | 65.8 | 100.7 | 39.4 | 52.5 | 68.5 | 58.1 | 12.55 |
| *23* | 72.6 | 92.3 | 41.4 | 54.1 | 65.0 | 78.1 | 14.13 |
| *24* | 98.2 | 129.9 | 47.7 | 74.1 | 106.5 | 62.5 | 13.77 |
| *1925* | 67.4 | 141.6 | 38.0 | 58.8 | 78.2 | 186.8 | 13.94 |
| *26* | 64.2 | 119.8 | 39.8 | 57.5 | 83.4 | 141.4 | 14.09 |
| *27* | 72.3 | 111.8 | 45.0 | 67.8 | 85.2 | 96.4 | 11.36 |

*Yearbook of Dep't of Agriculture, 1916, 1927.*

Table XVI.—Wheat: Estimated Price per Bushel Received by Producers, United States, on the 15th of the Month

| *Year* | *Jan.* | *Feb.* | *Mch.* | *Apr.* | *May* | *June* | *July* | *Aug.* | *Sept.* | *Oct.* | *Nov.* | *Dec.* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1914 | 118.8 | 131.8 | 132.6 | 135.6 | 135.6 | 117.2 | 76.7 | 84.9 | 93.4 | 95.4 | 97.9 | 103.2 |
| 15 | 108.4 | 108.4 | 100.8 | 100.6 | 101.2 | 96.5 | 104.6 | 100.8 | 93.0 | 92.0 | 92.5 | 97.4 |
| 16 | 157.6 | 164.6 | 172.2 | 213.0 | 247.2 | 234.3 | 100.0 | 119.2 | 133.8 | 147.4 | 159.4 | 155.3 |
| 17 | 201.6 | 202.0 | 202.6 | 203.1 | 203.0 | 202.8 | 224.5 | 219.3 | 205.2 | 200.3 | 200.4 | 201.4 |
| 18 | 206.2 | 207.8 | 211.1 | 222.6 | 229.8 | 225.2 | 203.8 | 205.0 | 205.7 | 205.9 | 205.1 | 204.5 |
| 19 | 233.8 | 231.2 | 230.3 | 242.6 | 250.8 | 256.0 | 219.6 | 211.4 | 207.6 | 211.4 | 214.0 | 223.4 |
| 1920 | 149.2 | 148.2 | 140.4 | 122.1 | 119.0 | 119.8 | 242.9 | 225.4 | 216.5 | 201.2 | 165.8 | 146.4 |
| 21 | 95.2 | 107.0 | 117.0 | 119.0 | 118.8 | 109.6 | 108.5 | 103.0 | 103.4 | 99.9 | 93.4 | 93.0 |
| 22 | 104.6 | 104.4 | 106.0 | 108.4 | 108.2 | 100.8 | 99.8 | 92.6 | 89.2 | 94.1 | 99.4 | 103.2 |
| 23 | 96.7 | 98.0 | 98.8 | 95.8 | 96.8 | 98.5 | 89.6 | 86.4 | 91.0 | 94.2 | 93.7 | 94.5 |
| 24 | 162.1 | 169.8 | 164.0 | 140.5 | 149.1 | 152.7 | 105.8 | 116.8 | 114.2 | 129.7 | 133.6 | 141.1 |

*Yearbook of the Dep't of Agriculture, 1925, p. 764.*

Table XVII.—Index Numbers of Wholesale Prices by Groups and Sub-Groups, by Months, 1924.

| MONTH | *Grains* | *Live Stock and Poultry* | *Other Farm Products* | *All Farm Products* |
|---|---|---|---|---|
| Jan. | 121.0 | 103.8 | 193.9 | 144.4 |
| Feb. | 123.7 | 105.2 | 185.8 | 143.0 |
| March | 120.7 | 110.4 | 170.1 | 137.2 |
| Apr. | 118.4 | 114.3 | 170.7 | 138.5 |
| May | 120.1 | 110.4 | 168.6 | 136.4 |
| June | 126.5 | 104.9 | 164.7 | 134.0 |
| July | 144.6 | 109.4 | 168.8 | 140.9 |
| Aug. | 150.3 | 118.0 | 168.3 | 145.3 |
| Sept. | 151.6 | 116.6 | 163.7 | 143.1 |
| Oct. | 162.4 | 123.5 | 166.6 | 149.2 |
| Nov. | 166.8 | 113.7 | 174.9 | 149.5 |
| Dec. | 185.5 | 118.5 | 178.7 | 156.7 |
| AV. FOR YR. | 141.3 | 112.4 | 173.6 | 143.4 |

*Bulletin No. 390, U. S. Bureau of Labor Statistics, p. 17.*

**Table XVIII.—Index Numbers of Wholesale Prices, by Groups of Commodities, 1924. (Base: Estimated Value in 1913 = 100.)**

| *Group* | *Index Number* | *Group* | *Index Number* |
|---|---|---|---|
| Farm products............ | 143.4 | Chemicals and drugs...... | 130.4 |
| Foods.................... | 144.2 | House furnishings........ | 172.8 |
| Cloths and clothing........ | 190.9 | Miscellaneous............ | 116.7 |
| Fuel and lighting.......... | 170.3 | | |
| Metals and metal products.. | 134.5 | | |
| Building materials......... | 175.1 | | |

*Bulletin No. 390, U. S. Bureau of Labor Statistics, pp. 8–9.*

**Table XIX.—Average Daily Output per Man for Ten Representative Mines in the Fairmont, West Virginia, Coal Fields.**

| *Range of Output in Tons per Man per Day* | *Number of Times This Output Was Secured* | *Range* | *Frequency* |
|---|---|---|---|
| 2.50–2.99 | 1 | 5.50–5.99 | 12 |
| 3.00–3.49 | 2 | 6.00–6.49 | 14 |
| 3.50–3.99 | 7 | 6.50–6.99 | 2 |
| 4.00–4.49 | 5 | 7.00–7.49 | 3 |
| 4.50–4.99 | 13 | 7.50–7.99 | 2 |
| 5.00–5.49 | 22 | 8.00–8.49 | 2 |
| | | Total............ | 85 |

*Bulletin No. 361, U. S. Bureau of Labor Statistics.*

**Table XX.—Deaths and Death-Rates on Unmarried Men in Scotland, 1863.**

| *Age* | *Number Living* | *Deaths* | *Death-rate* | *Age* | *Number Living* | *Deaths* | *Death-rate* |
|---|---|---|---|---|---|---|---|
| 20–25 | 106,587 | 1,251 | 11.7 | 60– 65 | 5,242 | 227 | 43.3 |
| 25–30 | 48,618 | 666 | 13.7 | 65– 70 | 2,848 | 156 | 54.8 |
| 30–35 | 25,962 | 383 | 14.8 | 70– 75 | 2,021 | 205 | 101.4 |
| 35–40 | 15,857 | 253 | 16.0 | 75– 80 | 1,081 | 157 | 145.4 |
| 40–45 | 12,311 | 208 | 16.9 | 80– 85 | 513 | 101 | 196.9 |
| 45–50 | 8,824 | 179 | 20.3 | 85– 90 | 151 | 32 | 211.9 |
| 50–55 | 7,636 | 205 | 26.8 | 90– 95 | 50 | 21 | 420.0 |
| 55–60 | 5,550 | 142 | 25.6 | 95–100 | 6 | 3 | 500.0 |

*Quarterly Publications of the American Statistical Association, Mar. 1914, p. 55.*

Table XXI —Squares and Cubes: Square Roots.

| No. | *Square* | *Cube* | *Square Root* | No. | *Square* | *Cube* | *Square Root* |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1.000 | 51 | 2,601 | 132,651 | 7.141 |
| 2 | 4 | 8 | 1.414 | 52 | 2,704 | 140,608 | 7.211 |
| 3 | 9 | 27 | 1.732 | 53 | 2,809 | 148,877 | 7.280 |
| 4 | 16 | 64 | 2.000 | 54 | 2,916 | 157,464 | 7.348 |
| 5 | 25 | 125 | 2.236 | 55 | 3,025 | 166,375 | 7.416 |
| 6 | 36 | 216 | 2.449 | 56 | 3,136 | 175,616 | 7.483 |
| 7 | 49 | 343 | 2.646 | 57 | 3,249 | 185,193 | 7.550 |
| 8 | 64 | 512 | 2.828 | 58 | 3,364 | 195,112 | 7.616 |
| 9 | 81 | 729 | 3.000 | 59 | 3,481 | 205,379 | 7.681 |
| 10 | 100 | 1,000 | 3.162 | 60 | 3,600 | 216,000 | 7.746 |
| 11 | 121 | 1,331 | 3.317 | 61 | 3,721 | 226,981 | 7.810 |
| 12 | 144 | 1,728 | 3.464 | 62 | 3,844 | 238,328 | 7.874 |
| 13 | 169 | 2,197 | 3.606 | 63 | 3,969 | 250,047 | 7.937 |
| 14 | 196 | 2,744 | 3.742 | 64 | 4,096 | 262,144 | 8.000 |
| 15 | 225 | 3,375 | 3.873 | 65 | 4,225 | 274,625 | 8.062 |
| 16 | 256 | 4,096 | 4.000 | 66 | 4,356 | 287,496 | 8.124 |
| 17 | 289 | 4,913 | 4.123 | 67 | 4,489 | 300,763 | 8.185 |
| 18 | 324 | 5,832 | 4.243 | 68 | 4,624 | 314,432 | 8.246 |
| 19 | 361 | 6,859 | 4.359 | 69 | 4,761 | 328,509 | 8.307 |
| 20 | 400 | 8,000 | 4.472 | 70 | 4,900 | 343,000 | 8.367 |
| 21 | 441 | 9,261 | 4.583 | 71 | 5,041 | 357,911 | 8.426 |
| 22 | 484 | 10,648 | 4.690 | 72 | 5,184 | 373,248 | 8.485 |
| 23 | 529 | 12,167 | 4.796 | 73 | 5,329 | 389,017 | 8.544 |
| 24 | 576 | 13,824 | 4.899 | 74 | 5,476 | 405,224 | 8.602 |
| 25 | 625 | 15,625 | 5.000 | 75 | 5,625 | 421,875 | 8.660 |
| 26 | 676 | 17,576 | 5.099 | 76 | 5,776 | 438,976 | 8.718 |
| 27 | 729 | 19,683 | 5.196 | 77 | 5,929 | 456,533 | 8.775 |
| 28 | 784 | 21,952 | 5.292 | 78 | 6,084 | 474,552 | 8.832 |
| 29 | 841 | 24,389 | 5.385 | 79 | 6,241 | 493,039 | 8.888 |
| 30 | 900 | 27,000 | 5.477 | 80 | 6,400 | 512,000 | 8.944 |
| 31 | 961 | 29,791 | 5.568 | 81 | 6,561 | 531,441 | 9.000 |
| 32 | 1,024 | 32,768 | 5.657 | 82 | 6,724 | 551,368 | 9.055 |
| 33 | 1,089 | 35,937 | 5.745 | 83 | 6,889 | 571,787 | 9.110 |
| 34 | 1,156 | 39,304 | 5.831 | 84 | 7,056 | 592,704 | 9.165 |
| 35 | 1,225 | 42,875 | 5.916 | 85 | 7,225 | 614,125 | 9.220 |
| 36 | 1,296 | 46,656 | 6.000 | 86 | 7,396 | 636,056 | 9.274 |
| 37 | 1,369 | 50,653 | 6.083 | 87 | 7,569 | 658,503 | 9.327 |
| 38 | 1,444 | 54,872 | 6.164 | 88 | 7,744 | 681,472 | 9.381 |
| 39 | 1,521 | 59,319 | 6.245 | 89 | 7,921 | 704,969 | 9.434 |
| 40 | 1,600 | 64,000 | 6.325 | 90 | 8,100 | 729,000 | 9.487 |
| 41 | 1,681 | 68,921 | 6.403 | 91 | 8,281 | 753,571 | 9.539 |
| 42 | 1,764 | 74,088 | 6.481 | 92 | 8,464 | 778,688 | 9.592 |
| 43 | 1,849 | 79,507 | 6.557 | 93 | 8,649 | 804,357 | 9.644 |
| 44 | 1,936 | 85,184 | 6.633 | 94 | 8,836 | 830,584 | 9.695 |
| 45 | 2,025 | 91,125 | 6.708 | 95 | 9,025 | 857,375 | 9.747 |
| 46 | 2,116 | 97,336 | 6.782 | 96 | 9,216 | 884,736 | 9.798 |
| 47 | 2,209 | 103,823 | 6.856 | 97 | 9,409 | 912,673 | 9.849 |
| 48 | 2,304 | 110,592 | 6.928 | 98 | 9,604 | 941,192 | 9.899 |
| 49 | 2,401 | 117,649 | 7.000 | 99 | 9,801 | 970,299 | 9.950 |
| 50 | 2,500 | 125,000 | 7.071 | 100 | 10,000 | 1,000,000 | 10.000 |

**Table XXII.—Logarithms.**

| No. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | .0000 | .0043 | .0086 | .0128 | .0170 | .0212 | .0253 | .0294 | .0334 | .0374 |
| .1 | 0414 | 0453 | 0492 | 0531 | 0569 | 0607 | 0645 | 0682 | 0719 | 0755 |
| .2 | 0792 | 0828 | 0864 | 0899 | 0934 | 0969 | 1004 | 1038 | 1072 | 1106 |
| .3 | 1139 | 1173 | 1206 | 1239 | 1271 | 1303 | 1335 | 1367 | 1399 | 1430 |
| .4 | 1461 | 1492 | 1523 | 1553 | 1584 | 1614 | 1644 | 1673 | 1703 | 1732 |
| 1.5 | .1761 | .1790 | .1818 | .1847 | .1875 | .1903 | .1931 | .1959 | .1987 | .2014 |
| .6 | 2041 | 2068 | 2095 | 2122 | 2148 | 2175 | 2201 | 2227 | 2253 | 2279 |
| .7 | 2304 | 2330 | 2355 | 2380 | 2405 | 2430 | 2455 | 2480 | 2504 | 2529 |
| .8 | 2553 | 2577 | 2601 | 2625 | 2648 | 2672 | 2695 | 2718 | 2742 | 2765 |
| .9 | 2788 | 2810 | 2833 | 2856 | 2878 | 2900 | 2923 | 2945 | 2967 | 2989 |
| 2.0 | .3010 | .3032 | .3054 | .3075 | .3096 | .3118 | .3139 | .3160 | .3181 | .3201 |
| .1 | 3222 | 3243 | 3263 | 3284 | 3304 | 3324 | 3345 | 3365 | 3385 | 3404 |
| .2 | 3424 | 3444 | 3464 | 3483 | 3502 | 3522 | 3541 | 3560 | 3579 | 3598 |
| .3 | 3617 | 3636 | 3655 | 3674 | 3692 | 3711 | 3729 | 3747 | 3766 | 3784 |
| .4 | 3802 | 3820 | 3838 | 3856 | 3874 | 3892 | 3909 | 3927 | 3945 | 3962 |
| 2.5 | .3979 | .3997 | .4014 | .4031 | .4048 | .4065 | .4082 | .4099 | .4116 | .4133 |
| .6 | 4150 | 4166 | 4183 | 4200 | 4216 | 4232 | 4249 | 4265 | 4281 | 4298 |
| .7 | 4314 | 4330 | 4346 | 4362 | 4378 | 4393 | 4409 | 4425 | 4440 | 4456 |
| .8 | 4472 | 4487 | 4502 | 4518 | 4533 | 4548 | 4564 | 4579 | 4594 | 4609 |
| .9 | 4624 | 4639 | 4654 | 4669 | 4683 | 4698 | 4713 | 4728 | 4742 | 4757 |
| 3.0 | .4771 | .4786 | .4800 | .4814 | .4829 | .4843 | .4857 | .4871 | .4886 | .4900 |
| .1 | 4914 | 4928 | 4942 | 4955 | 4969 | 4983 | 4997 | 5011 | 5024 | 5038 |
| .2 | 5051 | 5065 | 5079 | 5092 | 5105 | 5119 | 5132 | 5145 | 5159 | 5172 |
| .3 | 5185 | 5198 | 5211 | 5224 | 5237 | 5250 | 5263 | 5276 | 5289 | 5302 |
| .4 | 5315 | 5328 | 5340 | 5353 | 5366 | 5378 | 5391 | 5403 | 5416 | 5428 |
| 3.5 | .5441 | .5453 | .5465 | .5478 | .5490 | .5502 | .5514 | .5527 | .5539 | .5551 |
| .6 | 5563 | 5575 | 5587 | 5599 | 5611 | 5623 | 5635 | 5647 | 5658 | 5670 |
| .7 | 5682 | 5694 | 5705 | 5717 | 5729 | 5740 | 5752 | 5763 | 5775 | 5786 |
| .8 | 5798 | 5809 | 5821 | 5832 | 5843 | 5855 | 5866 | 5877 | 5888 | 5899 |
| .9 | 5911 | 5922 | 5933 | 5944 | 5955 | 5966 | 5977 | 5988 | 5999 | 6010 |
| 4.0 | .6021 | .6031 | .6042 | .6053 | .6064 | .6075 | .6085 | .6096 | .6107 | .6117 |
| .1 | 6128 | 6138 | 6149 | 6160 | 6170 | 6180 | 6191 | 6201 | 6212 | 6222 |
| .2 | 6232 | 6243 | 6253 | 6263 | 6274 | 6284 | 6294 | 6304 | 6314 | 6325 |
| .3 | 6335 | 6345 | 6355 | 6365 | 6375 | 6385 | 6395 | 6405 | 6415 | 6425 |
| .4 | 6435 | 6444 | 6454 | 6464 | 6474 | 6484 | 6493 | 6503 | 6513 | 6522 |
| 4.5 | .6532 | .6542 | .6551 | .6561 | .6571 | .6580 | .6590 | .6599 | .6609 | .6618 |
| .6 | 6628 | 6637 | 6646 | 6656 | 6665 | 6675 | 6684 | 6693 | 6702 | 6712 |
| .7 | 6721 | 6730 | 6739 | 6749 | 6758 | 6767 | 6776 | 6785 | 6794 | 6803 |
| .8 | 6812 | 6821 | 6830 | 6839 | 6848 | 6857 | 6866 | 6875 | 6884 | 6893 |
| .9 | 6902 | 6911 | 6920 | 6928 | 6937 | 6946 | 6955 | 6964 | 6972 | 6981 |

Table XXII Continued.—Logarithms.

| No. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 5.0 | .6990 | .6998 | .7007 | .7016 | .7024 | .7033 | .7042 | .7050 | .7059 | .7067 |
| .1 | 7076 | 7084 | 7093 | 7101 | 7110 | 7118 | 7126 | 7135 | 7143 | 7152 |
| .2 | 7160 | 7168 | 7177 | 7185 | 7193 | 7202 | 7210 | 7218 | 7226 | 7235 |
| .3 | 7243 | 7251 | 7259 | 7267 | 7275 | 7284 | 7292 | 7300 | 7308 | 7316 |
| .4. | 7324 | 7332 | 7340 | 7348 | 7356 | 7364 | 7372 | 7380 | 7388 | 7396 |
| 5.5 | .7404 | .7412 | .7419 | .7427 | .7435 | .7443 | .7451 | .7459 | .7466 | .7474 |
| .6 | 7482 | 7490 | 7497 | 7505 | 7513 | 7520 | 7528 | 7536 | 7543 | 7551 |
| .7 | 7559 | 7566 | 7574 | 7582 | 7589 | 7597 | 7604 | 7612 | 7619 | 7627 |
| .8 | 7634 | 7642 | 7649 | 7657 | 7664 | 7672 | 7679 | 7686 | 7694 | 7701 |
| .9 | 7709 | 7716 | 7723 | 7731 | 7738 | 7745 | 7752 | 7760 | 7767 | 7774 |
| 6.0 | .7782 | .7789 | .7796 | .7803 | .7810 | .7818 | .7825 | .7832 | .7839 | .7846 |
| .1 | 7853 | 7860 | 7868 | 7875 | 7882 | 7889 | 7896 | 7903 | 7910 | 7917 |
| .2 | 7924 | 7931 | 7938 | 7945 | 7952 | 7959 | 7966 | 7973 | 7980 | 7987 |
| .3 | 7993 | 8000 | 8007 | 8014 | 8021 | 8028 | 8035 | 8041 | 8048 | 8055 |
| .4 | 8062 | 8069 | 8075 | 8082 | 8089 | 8096 | 8102 | 8109 | 8116 | 8122 |
| 6.5 | .8129 | .8136 | .8142 | .8149 | .8156 | .8162 | .8169 | .8176 | .8182 | .8189 |
| .6 | 8195 | 8202 | 8209 | 8215 | 8222 | 8228 | 8235 | 8241 | 8248 | 8254 |
| .7 | 8261 | 8267 | 8274 | 8280 | 8287 | 8293 | 8299 | 8306 | 8312 | 8319 |
| .8 | 8325 | 8331 | 8338 | 8344 | 8351 | 8357 | 8363 | 8370 | 8376 | 8382 |
| .9 | 8388 | 8395 | 8401 | 8407 | 8414 | 8420 | 8426 | 8432 | 8439 | 8445 |
| 7.0 | .8451 | .8457 | .8463 | .8470 | .8476 | .8482 | .8488 | .8494 | .8500 | .8506 |
| .1 | 8513 | 8519 | 8525 | 8531 | 8537 | 8543 | 8549 | 8555 | 8561 | 8567 |
| .2 | 8573 | 8579 | 8585 | 8591 | 8597 | 8603 | 8609 | 8615 | 8621 | 8627 |
| .3 | 8633 | 8639 | 8645 | 8651 | 8657 | 8663 | 8669 | 8675 | 8681 | 8686 |
| .4 | 8692 | 8698 | 8704 | 8710 | 8716 | 8722 | 8727 | 8733 | 8739 | 8745 |
| 7.5 | .8751 | .8756 | .8762 | .8768 | .8774 | .8779 | .8785 | .8791 | .8797 | .8802 |
| .6 | 8808 | 8814 | 8820 | 8825 | 8831 | 8837 | 8842 | 8848 | 8854 | 8859 |
| .7 | 8865 | 8871 | 8876 | 8882 | 8887 | 8893 | 8899 | 8904 | 8910 | 8915 |
| .8 | 8921 | 8927 | 8932 | 8938 | 8943 | 8949 | 8954 | 8960 | 8965 | 8971 |
| .9 | 8976 | 8982 | 8987 | 8993 | 8998 | 9004 | 9009 | 9015 | 9020 | 9025 |
| 8.0 | .9031 | .9036 | .9042 | .9047 | .9053 | .9058 | .9063 | .9069 | .9074 | .9079 |
| .1 | 9085 | 9090 | 9096 | 9101 | 9106 | 9112 | 9117 | 9122 | 9128 | 9133 |
| .2 | 9138 | 9143 | 9149 | 9154 | 9159 | 9165 | 9170 | 9175 | 9180 | 9186 |
| .3 | 9191 | 9196 | 9201 | 9206 | 9212 | 9217 | 9222 | 9227 | 9232 | 9238 |
| .4 | 9243 | 9248 | 9253 | 9258 | 9263 | 9269 | 9274 | 9279 | 9284 | 9289 |
| 8.5 | .9294 | .9299 | .9304 | .9309 | .9315 | .9320 | .9325 | .9330 | .9335 | .9340 |
| .6 | 9345 | 9350 | 9355 | 9360 | 9365 | 9370 | 9375 | 9380 | 9385 | 9390 |
| .7 | 9395 | 9400 | 9405 | 9410 | 9415 | 9420 | 9425 | 9430 | 9435 | 9440 |
| .8 | 9445 | 9450 | 9455 | 9460 | 9465 | 9469 | 9474 | 9479 | 9484 | 9489 |
| .9 | 9494 | 9499 | 9504 | 9509 | 9513 | 9518 | 9523 | 9528 | 9533 | 9538 |
| 9.0 | .9542 | .9547 | .9552 | .9557 | .9562 | .9566 | .9571 | .9576 | .9581 | .9586 |
| .1 | 9590 | 9595 | 9600 | 9605 | 9609 | 9614 | 9619 | 9624 | 9628 | 9633 |
| .2 | 9638 | 9643 | 9647 | 9652 | 9657 | 9661 | 9666 | 9671 | 9675 | 9680 |
| .3 | 9685 | 9689 | 9694 | 9699 | 9703 | 9708 | 9713 | 9717 | 9722 | 9727 |
| .4 | 9731 | 9736 | 9741 | 9745 | 9750 | 9754 | 9759 | 9763 | 9768 | 9773 |
| 9.5 | .9777 | .9782 | .9786 | .9791 | .9795 | .9800 | .9805 | .9809 | .9814 | .9818 |
| .6 | 9823 | 9827 | 9832 | 9836 | 9841 | 9845 | 9850 | 9854 | 9859 | 9863 |
| .7 | 9868 | 9872 | 9877 | 9881 | 9886 | 9890 | 9894 | 9899 | 9903 | 9908 |
| .8 | 9912 | 9917 | 9921 | 9926 | 9930 | 9934 | 9939 | 9943 | 9948 | 9952 |
| .9 | 9956 | 9961 | 9965 | 9969 | 9974 | 9978 | 9983 | 9987 | 9991 | 9996 |

# BIBLIOGRAPHY

We present here a list of some of the more useful texts on Statistics. More extensive bibliographies will be found in the texts of Jerome, Kelley, Keynes, and Rietz listed below. Fundamental papers on the subject will be found in such periodicals as: the *Journal of the American Statistical Association*, the *Review of Economic Statistics*, the *Journal of the Royal Statistical Society*, and *Biometrika*.

Alexander, C., "School Statistics and Publicity," Silver, Burdett & Co., N. Y., 1919.

Babson, R. W., "Business Barometers," Wellesley Hills, Mass., 16th ed., 1923.

Bailey, W. B., and Cummings, John, "Statistics," A. C. McClurg, Chicago, 1917.

Bowley, A. L., "Elements of Statistics," 4th ed., Scribner's, N. Y., 1920.

Brinton, W. C., "Graphic Methods for Presenting Facts," Engineering Magazine Co., N. Y., 1914.

Chaddock, R. E., "Principles and Methods of Statistics," Houghton Mifflin, Boston, 1925.

Chambers, G. G., "An Introduction to Statistical Analysis," Crofts, N. Y., 1925.

Crum, W. L., and Patton, A. C., "Economic Statistics," A. W. Shaw, N. Y., 1925.

Davenport, C. B., "Statistical Methods, with Special Reference to Biological Variation," 3d ed., Wiley, N. Y., 1914.

Davies, G. R., "Introduction to Economic Statistics," Century, N. Y., 1922.

Day, E. E., "Statistical Analysis," Macmillan, N. Y., 1925.

Elderton, W. P., "Frequency Curves and Correlation," C. and E. Layton, London, 1906.

Fisher, Arne, "The Mathematical Theory of Probabilities," Macmillan, N. Y., 1922. "An Elementary Treatise on Frequency Curves," Macmillan, N. Y., 1922.

Fisher, Irving, "The Making of Index Numbers," Houghton Mifflin, Boston, 1922.

Forsyth, C. H., "Mathematical Analysis of Statistics," Wiley, N. Y., 1924.

Gavett, G. I., "Statistical Method," McGraw-Hill, N. Y., 1925.

Haskell, S. C., "How to Make and Use Graphic Charts," Codex Book Co., N. Y., 1923. "Graphic Charts in Business," Codex Book Co., N. Y., 1922.

Jerome, Harry, "Statistical Method," Harper's, N. Y., 1924.

Jones, D. C., "A First Course in Statistics," G. Bell and Son, London, 1921.

Jordan, D. F., "Business Forecasting," Prentice-Hall, N. Y., 1921.

Karsten, K. G., "Charts and Graphs," Prentice-Hall, N. Y., 1923.

Kelley, T. L., "Statistical Method," Macmillan, N. Y., 1923.

Keynes, J. M., "A Treatise on Probability," Macmillan, N. Y., 1921.

King, W. I., "Elements of Statistical Method," Macmillan, N. Y., 1913. "Wealth and Income of the People of the United States," Macmillan, N. Y., 1915.

Koren, John, "History of Statistics," Macmillan, N. Y., 1918.

Lipka, J., "Graphical and Mechanical Computation;" Wiley, N. Y., 1918.

Marshall, W. C., "Graphical Methods," McGraw-Hill, N. Y., 1921.

Mills, F. C., "Statistical Methods," Holt, N. Y., 1924.

Mitchell, W. C., "Business Cycles," University of California Press, 1913. "Index Numbers of Wholesale Prices in the United States and Foreign Countries," *Bureau of Labor Statistics Bulletin*, No. 284, Washington D. C., 1921.

Moore, H. L., "Economic Cycles: Their Law and Cause," Macmillan, N. Y., 1914. "Forecasting the Yield and the Price of Cotton," Macmillan, N. Y., 1917. "Generating Economic Cycles," Macmillan, N. Y., 1923. "Laws of Wages," Macmillan, 1911.

National Bureau of Economic Research, Inc., New York City. "Income in the United States—Its Amount and Distribution, 1909–1919."

Vol. I. General Summary of Findings.

Vol. II. Detailed Report.

Vol. III. Distribution of Income by States.

Vol. IV. Business Cycles and Unemployment, 1923.

Vol. V. W. I. King, "Employment, Hours, and Earnings in Prosperity and Depression—United States, 1920–22, 1923."

Pearl, R., "Medical Biometry and Statistics," W. B. Saunders, Philadelphia, 1923.

Persons, W. M., "Index of General Business Conditions," Harvard University Press, Cambridge Mass., 1919.

Persons, W. M., Foster, W. T., and Hettinger, A. J., (Editors) "The Problem of Business Forecasting," Houghton Mifflin, Boston, 1924.

Riegel, R., "Elements of Business Statistics," Appleton, N. Y., 1924.

Rietz, H. L., (Editor) "Handbook of Mathematical Statistics," Houghton Mifflin, Boston, 1924.

Rugg, H. O., "Statistical Methods Applied to Education," Houghton Mifflin, Boston, 1917.

Running, T. R., "Empirical Formulas," Wiley, N. Y., 1917.

Secrist, H., "An Introduction to Statistical Methods," Revised edition, Macmillan, N. Y., 1925. "Readings and Problems in Statistical Methods," Macmillan, N. Y., 1920.

Thorndike, E. L., "An Introduction to the Theory of Mental and Social Measurements," Columbia Univ., N. Y., 1915.

Walsh, C. M., "The Measurement of General Exchange Value," Macmillan, N. Y., 1901. "The Problem of Estimation," P. S. King and Son, London, 1921.

West, C. J., "Introduction to Mathematical Statistics," R. G. Adams, Columbus, 1918.

Whipple, G. C., "Vital Statistics," Wiley, N. Y., 1922.

Young, B. F., "Statistics as Applied in Business," Ronald Press, N. Y., 1925.

Yule, G. U., "An Introduction to the Theory of Statistics," 7th ed., C. Griffin and Co., London, 1924.

# INDEX

(The references are to pages)

D7086